UNIVERSITÀ DEGLI STUDI FIRENZE

Maastricht University

DOCTOR OF PHILOSOPHY IN LAW

European and Transnational Legal Studies

Cycle XXXV

ACADEMIC DISCIPLINE (SSD) IUS/17

Criminal Behavior and Accountability of Artificial Intelligence Systems

Doctoral Candidate

Dr. Alice Giannini

Supervisors

Prof Michele Papa

Prof. André Klip

Coordinator

Prof. Alessandro Simoni

Years 2019/2023

CRIMINAL BEHAVIOR AND ACCOUNTABILITY OF ARTIFICIAL INTELLIGENCE SYSTEMS

Dissertation to obtain the degree of Doctor
at the University of Florence and at Maastricht University

by

ALICE GIANNINI

Supervisors:

Prof. dr. A.H. Klip
Prof. dr. M. Papa

Assessment committee:
Prof. dr. S. van der Aa (Chair)
Prof. dr. D. Roef
Prof. dr. S. Pietropaoli, University of Florence
Prof. dr. A. Simoni, University of Florence
Prof. dr. J. Lelieur, University of Strasbourg
Prof. dr. L. Picotti, University of Verona

"Tutto questo avviene non sul mare, non nel sole, – pensa il nuotatore Palomar,
– ma dentro la mia testa, nei circuiti tra gli occhi e il cervello.
Sto nuotando nella mia mente;
è solo là che esiste questa spada di luce;
e ciò che mi attira è proprio questo.
È questo il mio elemento, l'unico che io possa in qualche modo conoscere."

Italo Calvino, Palomar (1984)

Tutta la vita, a provare a dirti che partivo
O che partivo o che morivo

Lucio Dalla, Tutta la vita (1984)

TABLE OF CONTENTS

ESSENTIAL GLOSSARY OF TERMS IV

1 INTRODUCTION 1

1.1 Science-Fiction, Legal-Fictions and the "Plasticity" of Criminal Law 1
1.2 The Definitional Problem of Artificial Intelligence 6
1.3 A Three-Ring Circus: E-Personhood, Failures of Causation, and Guilty Robot Minds 10
1.3.1 E-Personhood 10
1.3.2 Failures of Causation 13
1.3.3 Guilty Robot Minds 17
1.4 Research Question & Structure of Chapters 18
1.5 Methodology 20
1.6 AIs Going Bad – Examples 23
EXAMPLE A – THE DRIVERLESS CAR 23
EXAMPLE B – PREDICTING SUICIDE 23
EXAMPLE C – EVIL AI 24
EXAMPLE D – MORAL DILEMMAS 24

2 DEFINING ARTIFICIAL INTELLIGENCE 26

2.1 Introduction: to Define or not to Define? 26
2.2 A Short History of AI 30
2.3 Gaining a Basic Understanding of Modern AI Systems 35
2.4 Definitions of AI: an Overview 42
2.5 Adopted Working Definition 45
2.6 Conclusions 46

3 EXPANSIONISTS, MODERATES, SKEPTICS: THE SCHOLARLY DEBATE ON AI AND CRIMINAL LAW 48

3.1 Introduction. The scholarly debate on criminal liability of AI-systems 48
3.2 Expansionists (or the Front of Robotic Liberation) 51
3.2.1 Gabriel Hallevy – the AI "Believer" 51
3.2.2 Ying Hu: a Criminal Code for Robots 59
3.2.3 Christina Mulligan: Revenge Against Robots 64
3.2.4 Lasse Quarck: the German Exception 66
3.3 Moderates 68
3.3.1 Ryan Abbott and Alex Sarch: a General Theory for AI-Punishment 68
3.3.2 Lagioia and Sartor 75
3.3.3 Freitas, Andrade, Novais - Nora Osmani 81
3.3.4 Simmler and Markwalder 84
3.3.5 Mihail Diamantis: the Corporate Mind and Body Approach 88
3.4 Skeptics 89

3.4.1 The Italian Approach 89
3.4.2 Ugo Pagallo 96
3.4.3 Dafni Lima 102
3.4.4 Peter Asaro: A Body to Kick, but Still No Soul to Damn 105
3.4.5 The German Approach 107
3.4.5.1 Sabine Gless, Thomas Weigend, and Emily Silverman 108
3.4.5.2 Susanne Beck 113
3.4.5.3 Gerhard Seher 115
3.5 Conclusions 118

4 ASCRIPTION 122

4.1 Introduction. On Paperclips, Planes, and AI 122
4.2 Theories of Criminalization 126
4.3 Methodology and Structure of Chapters 5 and 6 128

5 CRIMINAL CAPACITY 130

5.1 What is an AI Agent, Exactly? 130
5.2 Capacity in Criminal Law 133
5.3 Artificial Intelligence Systems as Rechtspersonen? 138
5.3.1 A quick glimpse into the vexata quaestio of e-personhood 138
5.3.2 Criminal or Moral Machines? 142
5.3.2.1 The MIT's Moral Machine 145
5.3.2.2 The Burning Room Dilemma 147
5.3.2.3 Comment 149
5.3.3 Action control 152
5.4 On Artificial Insane and Infant Offenders 154
5.4.1 Artificial Insane Offenders 156
5.4.2 Artficial Infant Offenders 159
5.5 Preliminary conclusions 161

6 ARTIFICIAL INTELLIGENCE CRIME 165

6.1 Introduction: Glimmers of an Economic Theory of AI-Crime 165
6.2 Matters of mens rea 172
6.2.1 Overview 172
6.2.2 Responsibility of Machines 173
6.2.3 Responsibility of Humans 176
6.2.3.1 The DNA of Negligence 176
6.2.3.2 Human Oversight and Human in The Loop: Begging the Question? 178
6.2.3.3 Negligence Failures 180
6.2.3.4 It's All about the Data 186
6.3 Matters of actus reus 187
6.3.1 The Act 188
6.3.2 Failures of Causation 189

6.3.2.1 The "Many Hands Problem" 189
6.3.2.2 The Black Box Problem 190
6.3.2.3 The Shortcuts Problem 191
6.3.2.4 Omissions to Act 192
6.4 Corporate Criminal Liability for Automated decisions 194
6.4.1 Models of Corporate Criminal Liability ("CCL") 194
6.4.2 The Next Frontier? Diamantis' theory of Corporate Criminal Liability for Automated Decisions 198
6.5 Preliminary Conclusions 202

7 OVERVIEW OF EXISTING LEGAL FRAMEWORKS ON AI CRIMINAL LIABILITY 206

7.1 Introduction 206
7.2 General-Scope Tools 209
A) Council of Europe 209
B) Singapore 213
7.3 Self-Driving Tragedies: AV-Specific Tools 221
C) France 225
D) England and Scotland 227
E) Germany 233
7.4 Conclusions 241

8 CONCLUSIONS 243

8.1 Incipit 243
8.2 The Definitional Question 245
8.3 A mare magnum of Scholarly Literature. The Introspective Question 248
8.4 Anthropocentrism and Responsibility of Machines. The Attribution Question 252
8.4.1 Holding AI to a Higher (Moral) Standard 252
8.4.1.2 The (Ir)Relevance of Motives 255
8.4.2 Retributivists at Heart? 260
8.4.3 Re-Evaluating the Comparison Between AI and Corporations 262
8.4.4 Epicenters of Liability: the Human Culprit 266
8.5 General Conclusions. The Future- and Backward-Facing Question 269
8.5.1 Looking Forward 272
8.5.1 Looking Backwards 275

SUMMARY 278
SINTESI 281
SAMENVATTING 284
IMPACT STATEMENT 287
INDEX OF FIGURES 289
BIBLIOGRAPHY 290
BIOGRAPHY 314

ESSENTIAL GLOSSARY OF TERMS

Algorithm
Set of rules that define a sequence of operations to solve a problem.

Artificial intelligence
Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal.
AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions.
As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).[1]

Autonomous vehicles (AVs)
Vehicles capable of navigating streets and interpreting traffic without a driver actively operating. They can display six levels of automation:
0 – No Automation
1 – Driver Assistance
2 – Partial Driving Automation
3 – Conditional Driving Automation
4 – High Driving Automation
5 – Full Driving Automation[2]

Algorithm
Set of rules that precisely define a sequence of operations to solve a problem.

Artificial Neural Networks (ANNs)
Neural networks are made from multiple layers of artificial neurons encoded in software. Each neuron can be connected to others in the layers above. One neuron

[1] The definition was elaborated by the High-Level Expert Group on Artificial Intelligence (AI-HLEG), *A definition of AI: Main capabilities and Scientific Disciplines*, 2018. Available at: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf.

[2] Society for Automotive Engineers International (SAE), *J3016 Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, April 2021.

receives an 'input' (for example information on a pixel in a picture) and another neuron produces an 'output' (for example the classification of the picture). This technique is inspired by the functioning of the human brain. The knowledge of the ANN is stored in the network itself.

Automated decision system (ADS)
A computational process, including one derived from artificial intelligence techniques, which makes a decision without human involvement.

Bots
Automated program which runs on the Internet and typically mimics human behavior.

Computer vision
Process by which the system extracts data from a series of images or videos.

Deep Learning (DL)
Sub-set of ML where the system consists of layers of artificial neural networks (ANNs). The network analyzes data and identifies relevant features by itself.

Deep Neural Networks (DNNs)
DNNs are built on layers of interconnected artificial neurons which work together. A DNN will make decisions using an intuitive decision-making process and it will learn through experience.

Expert systems
Computer programs which resemble the decision making of an expert to solve problems by using inference procedures. Expert systems are based on symbolic AI (see definition below).

Generative Adversarial Network (GAN)
Type of ANN that learns by competing against each other. Usually one ANN (the generator) produces content and the other (the discriminator) detects whether it was produced by an ANN or by a human agent or if the output has any flaws. Each side improves from such interaction.

Human in The Loop (HITL)
All the operations are executed by the system under human control.

Human on The Loop (HOTL)
The system acts autonomously but under the supervision of a human.

Human post loop
The human agent cannot directly affect the operation of the system but it can intervene after the operation to block it

Human out of the loop
The system operates in absolute autonomy without any control, before or after the operation.

Machine Learning (ML)
AI technique where the algorithm learns to find on its own a solution to a set problem or task.

Natural Language Processing (NLP)
Process by which the system extracts data from human language and makes decisions based on that information. It enables clear human-to-machine communication. Examples of NLP systems are voice-activated digital assistants such as Alexa, Siri, Cortana and Google Assistant.

Reinforcement learning
Type of ML training process where the algorithm interacts with the environments and learns by trial, discovering errors or rewards.

Strong Artificial Intelligence or Artificial General Intelligence (AGI)
AI systems which display human cognitive abilities and can perform all sorts of tasks, learning from their experience. As of today, there are no AI systems which display general intelligence.

Symbolic AI
Also referred to as "expert systems" or "logical AI", It refers to an approach to model AI systems based on "if *x*, then *y*" rules. IBM's Deep Blue program, which beat Gary Kasparov at chess in 1997 is an example of symbolic AI.

Supervised learning
Type of ML where the programmer trains the system by defining a set of expected results for a selected set of input and provides the system with an evaluation of the results. The ML training process is guided by labelled data. This system generates hypothesis on how to classify the relevant traits. Every time the system errs the hypotheses are corrected. Supervised learning techniques include linear and logistic regression, Decision Trees, Naïve Bayes, etc.

Unsupervised learning
Type of ML training process where little to nothing labelled data is used and with low human intervention. The system will try to find connections and structure in the data (for example through clustering, i.e. creating groups of similar data objects) by extracting and analyzing useful features. It learns by its own observations and is highly dependent on the experience.

Weak Artificial Intelligence or Narrow Artificial Intelligence
AI systems which can perform only specific pre-set tasks and therefore operate in a limited domain. Current AI systems belong to this category.

1 INTRODUCTION

Rachael: It seems you feel our work is not a benefit to the public.
Deckard: Replicants are like any other machine - they're either a benefit or a hazard.
If they're a benefit, it's not my problem.
Blade Runner (1982)

1.1 Science-Fiction, Legal-Fictions and the "Plasticity" of Criminal Law – **1.2** The Definitional Problem of Artificial Intelligence– **1.3** A Three-Ring Circus: E-Personhood, Failures of Causation and Guilty Robot Minds – **1.3.1** E-Personhood – **1.3.2** Failures of Causation – **1.3.3** Guilty Robot Minds – **1.4.** Punishing Artificial Intelligence – **1.5.** Research Question & Structure of Chapters – **1.6.** Methodology

1.1 SCIENCE-FICTION, LEGAL-FICTIONS AND THE "PLASTICITY" OF CRIMINAL LAW[3]

"Intelligent systems currently cause real world harms without a collective memory of their failings".[4] This is the incipit of the research paper presenting the world's first systematized Artificial Intelligence Incidents Database (AIID).[5] The AIID is an industrial/non-profit cooperative collection of "intelligent system failures experienced in the real world",[6] i.e., incidents.[7] It aims at answering the question "what can go wrong when someone deploys this

[3] The title takes inspiration from the work of R. Abbott & A. Sarch, "Punishing Artificial Intelligence: Legal Fiction or Science Fiction", *UCD L Rev*, vo. 53, 2019, p. 323 and of K. Burchard, "Künstliche Intelligenz als Ende des Strafrechts? Zur algorithmischen Transformation der Gesellschaft", *Normative Orders Working Paper*, No. 2, 2019, p. 4.
[4] S. McGregor, "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database", *Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21),* 2021, p. 1.
[5] Available at: https://incidentdatabase.ai/.
[6] McGregor, 2021, p. 1.
[7] Cfr. Ch. 4. 1.

system?".[8] The existence of the AIID and other databases[9] tells us something: they give foundation with regards to the relevance of this research. They push us to ask ourselves a crucial question: if "something goes wrong" with an artificial intelligence (AI) system, should criminal law care? If yes, how? As it will be shown, these two (apparently) simple questions will be the tenet of this research and will be reformulated in terms of *necessity* and *feasibility* in the main research question.

Certainly, the idea of criminal behavior of AI systems is nothing new. In fact, science-fiction has been dealing for decades with "evil robots" rebelling against humans and taking control, or with machines that "go crazy" and act unpredictably:[10] one could actually argue that Asimov's three laws of robotics were nothing but the most famous attempt at regulating forms of AI misconduct.[11-12] Thus, differently from the most famous sci-fi chronicles, in the last decade AI has actually become part of our daily lives and is here to stay.

The most recent evolution in AI techniques has led to the development of systems capable of unsupervised, unforeseen, and autonomous actions. Algorithms, through machine learning (ML) techniques, can learn from their past actions and teach themselves new

---

[8] McGregor, 2021, p. 1.

[9] See also the AIAAIC Repository, an independent collection of over 850 incidents and controversies driven by and relating to AI, algorithms, and automation. Available at: https://docs.google.com/spreadsheets/d/1Bn55B4xz21-_Rgdr8BBb2lt0n_4rzLGxFADMlVW0PYI/edit#gid=1051812323.

[10] "Think instead of the false Maria in Metropolis (1927); Hal 9000 in 2001: A Space Odyssey (1968[…]; C3PO in Star Wars (1977); Rachael in Blade Runner (1982); Data in Star Trek: The Next Generation (1987); Agent Smith in The Matrix (1999) or the disembodied Samantha in Her (2013)", L. Floridi, Should we be afraid of AI?, *aeon.co*. Available at: https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible.

[11] "(1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law." I. Asimov, *I, Robot*, Harper Collins Publisher, 1950.

[12] S. N. Lehman-Wilzing, "Frankenstein unbound: Towards a legal definition of artificial intelligence", *Futures*, Vol. 13, Issue 6, 1981, p. 445.

patterns of behavior.[13] Such methods, then, possibly allow for algorithmic misbehavior without any human intervention.[14]

As stated by the European Parliament in its resolution containing recommendations to the Commission on a *civil* liability regime for AI, "[…] the opacity, connectivity and autonomy of AI-systems could make it in practice very difficult or even impossible to trace back specific harmful actions of AI-systems to specific human input or to decisions in the design".[15] Undeniably, "[a]s a transformative technology that is characterised by high complexity, unpredictability and autonomy in its decision-making and learning capacities, *AI has the potential to challenge traditional notions of legal personality, individual agency and responsibility"*.[16]

At this point, it is clear that things are on the move: scholars and policy makers have started asking themselves questions on how to fill this apparent gap – not just in the field of civil liability. This will be evident in Chapter 3, which studies and classifies the scholarly debate on AI and criminal liability, where instead policymaking initiatives are the object of Chapter 5. One relevant example in the field of criminal law which is worth of mention is the work conducted by the Council of Europe Committee on Criminal Problems (CDPC). In September 2020, the CDPC published a feasibility study on a future Council of Europe (CoE) instrument on artificial intelligence and criminal law. During a thematic session on

---

[13] "A few features of AI are important to highlight. First, AI has the potential to act unpredictably. Some leading AIs rely on machine learning or similar technologies which involve a computer program, initially created by individuals, further developing in response to data without explicit programming. This is one means by which AI can engage in activities its original programmers may not have intended or foreseen. Second, AI has the potential to act unexplainably. It may be possible to determine what an AI has done, but not how or why it acted as it did. This has led to some AIs being described as "black box" systems. […] That is particularly likely in the case of AIs that learn from data, and which may have been exposed to millions or billions of data points. Even if it is theoretically possible to explain an AI outcome, it may be impracticable given the potentially resource intensive nature of such inquiries, and the need to maintain earlier iterative versions of AI and specific data. Third, AI may act autonomously", Abbott & Sarch, 2019, pp. 330-331.

[14] "For our purposes, that is to say an AI may cause harm without being directly controlled by an individual". Ivi, p. 331.

[15] European Parliament, *Resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence*, (2020/2014(INL)), 20 October 2020.

[16] M. Kritikos, European Parliamentary Research Service – Scientific Foresight Unit, Briefing "Artificial Intelligence ante portas: Legal & ethical reflections", 2019, p. 1.

AI and criminal law responsibility, focused on the case of automated driving,[17] prof. Sabine Gless, Special Rapporteur, underlined that

> Existing liability models may be inadequate to address the future role of AI in criminal activities. This may undermine the certainty of the law. It may leave specific gaps given the nature of AI that is based on machine learning and gives space to a machine actor whose action we do not entirely understand.[18]

The initiative of the CPDC is just the tip of the iceberg: for example, the broader topic of "Artificial Intelligence and Criminal Justice" was picked by the International Association of Penal Law for the XXI International Congress of Penal Law of 2024.[19] The trend will be analyzed in depth in the thesis.

As it will be shown, modern legal systems are facing situations in which it will be hard – if not impossible – to determine with a legally acceptable degree of certainty whether the harm caused by an AI system on a protected legal interest can be ascribed to a *human* agent involved in the causal chain of events (be it the programmer, the producer, or the user of the technology).[20] In other words, AI systems stretch the distance between the harmful event

---

[17] Council of Europe – European Committee on Crime Problems (CDPC), Concept Paper "Artificial intelligence and criminal law responsibility in Council of Europe member states - the case of automated vehicles", CDPC(2018)14Rev, 16 October 2018. Available at: https://rm.coe.int/cdpc-2018-14rev-artificial-intelligence-and-criminal-law-project-2018-/16808e64ad.

[18] Council of Europe – CDPC, Thematic session on Artificial Intelligence and Criminal Law of 28 November 2018, Final remarks by Professor Sabine Gless, Special Rapporteur, "Artificial intelligence and its impact on CDPC work. The case of automated driving", cdpc/docs 2018/cdpc (2018)22, 28 November 2018.

[19] See the concept paper by K. Ligeti, *Artificial Intelligence and Criminal Justice*. Available at: https://www.penal.org/sites/default/files/Concept%20Paper_AI%20and%20Criminal%20Justice_Ligeti.pdf. See also the questionnaire elaborated by L. Picotti for the AIDP, Section I "Traditional Criminal Law Categories and AI: Crisis or Palingenesis?". Available at: https://www.penal.org/sites/default/files/Questionnaires%20EN.pdf.

[20] "*Eine besondere Rolle mit Blick auf die strafrechtliche Verantwortung nehmen Maschinen ein, die einen eigenen Entscheidungsspielraum haben, durch Sensoren und Vernetzung Informationen erhalten und selbst auswerten. In diesen Fällen […] lässt sich weder im Vorhinein vorhersehen, welche Entscheidungen die Maschinen in welchen Situationen treffen werden, noch im Nachhinein feststellen, worauf die Entscheidungen beruhten. Insbesondere ob einer der Beteiligten, d.h. der Programmierer, Produzent oder der Nutzer einen Fehler gemacht hat, ist häufig nicht mehr nachweisbar*", S. Beck, "Die Diffusion strafrechtlicher Verantwortlichkeit durch Digitalisierung und Lernende Systeme", *Zeitschrift für Internationale Strafrechtsdogmatik*, Vol. 2, 2020, p. 44.

and the responsible human person.[21] In this sense, one can speak about a *liability gap*. Machines are "inducing some problems that are specific to criminal law" and, consequently, "[…] we have to determine whether the behavior of robots falls within the loopholes of the system, necessitating the intervention of lawmakers at both national and international levels".[22]

It can be affirmed that AI systems today represent a sort of "*stress test*" for traditional human-based criminal liability frameworks. It is in this perspective that the question of ascribing criminal liability *to* AI systems – "AIs going bad" - takes on new relevance.[23]

The main research question which will be guiding this research is: *to what extent is a theoretical framework of criminal law for liability of non-human agents needed and feasible?* This study will explore cases in which a crime is functionally committed by a machine and it is "not practically defensible"[24] to impute the AI's behaviors to "bad" human actors and cases where there is no other identifiable human agent that acted with criminal culpability. Such an investigation will entail a discussion on what can be attributed to an AI system, to humans, or to both.

A discussion on the attribution of criminal liability to AI systems is necessarily connected to *purposes* of criminal punishment. As such, this study necessarily will touch upon different theories of punishment. Moreover, the criminalization of conducts is usually guided by certain criteria which are selected according to theories of criminalization. In other words, criminal law determines what kinds of "undesirable conducts"[25] should be punished and it does so in pursuance of different aims. Consequently, this study will inevitably discuss theories of criminalization.

In conclusion, we will embark on this research conscious of, and strong of, the *plasticity* of criminal law theory. As it was argued,

---

[21] M. Hildebrandt, "Technology", in M. Dubber & T. Hörnle (Eds.), *The Oxford Handbook of Criminal Law*, Oxford University Press, 2014, p.190.

[22] Pagallo, *The Laws of Robots: Crimes, Contracts and Torts*, Springer, 2013, p. 45.

[23] As stated by Burchard, "Soweit KI strafrechtswissenschaftlich bereits in den Blick genommen wird, wird sie […] herkömmlich als potentiell *regulierungsbedürftiger* und auch *regulierungs-fähiger* Objektbereich geführt". Burchard, 2019, p. 4.

[24] Abbott & Sarch, 2019, p. 328.

[25] W. R. LaFave, *Criminal Law*, 6th Edition, West Academic, 2017, p. 26.

> The use of legal fictions to solve difficult conceptual questions or practical problems - such as how to conceptualize or prove particular sorts of mental elements for AI or misbehavior by its developers - gives criminal law theory impressive *plasticity*.
> Legal fictions help turn the criminal law into a pragmatic tool for solving social problems.[26]

It is necessary to commence the inquiry with a critical eye. Indeed, criminal law is not the only tool through which society can obtain a desirable conduct: think of example of civil remedies or of administrative sanctions. This entails distancing ourselves from ambitions of supremacy of the criminal sanction, which could eventually turn out to be unsuited for the case at hand, and to which we tend to attribute a symbolic-restorative function that perhaps is not quite its.

1.2 THE DEFINITIONAL PROBLEM OF ARTIFICIAL INTELLIGENCE

There is no overall accepted definition of AI.[27-28] Hence, it will be crucial to address the issue in order to avoid fallacies throughout the research.

The term "Artificial Intelligence" was first coined at the famous 1956 Dartmouth College meeting by John McCarthy[29] and is used by scientists today to refer to a research field which includes a variety of techniques and technologies such as "theorem proving, heuristic search, game playing, expert systems, neural networks, Bayesian networks, data

---

[26] Ivi, p. 384.

[27] For a systematic analysis of the issue of defining "artificial intelligence" see inter alia: P. Wang, "On Defining Artificial Intelligence", *Journal of Artificial General Intelligence*, Vol. 10, No. 2, 2019, pp. 1-37.

[28] For an overview of what AI is in relation to the practice of law, see H. Surden, "Artificial Intelligence and Law: an Overview", *Ga. St. U. L. Rev.*, Vol. 35, 2019. For arguments in favor of policy makers using terms other than "artificial intelligence" for regulatory purposes see J. Schuett, "A Legal Definition of AI", *Xiv:1909.01095 [cs.CY]*, 2019.

[29] "[T]he artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving". J. McCarthy et al., *A Proposal for the Dart-mouth Summer Research Project on Artificial Intelligence*, 1955. Available at: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

mining, agents, and recently, deep learning".[30] In a problem-solving perspective, and in its most comprehensive sense, AI technology aims at automating tasks that "are thought to involve intelligence when people perform them […]"[31]. Differently put, "[w]hen engineers automate an activity that requires cognitive activity when performed by humans, it is common to describe this as an application of AI".[32]

The biggest difficulty when defining an AI agent is, of course, embedded in the definition of (human) intelligence itself. Nevertheless, this study will not embark on an inquiry on what intelligence is or is not, nor on one regarding whether "artificial" intelligence constitutes "real" intelligence or not.[33] However, selected concepts which pertain to such philosophical debate – such as the notion of free will and autonomy – will be addressed when discussing criminal capacity, *mens rea*, and *actus reus* questions. In other words, the framework of reference for the analysis will always be the classical construction of criminal responsibility.

AI systems deployed today can be defined as "weak" or "specific" AI systems, as opposed to "strong" or "general" AI systems. Weak AI Systems are capable of executing pre-set tasks – "often at a level above human capabilities",[34] where instead Artificial General Intelligence aims at realizing systems capable of exhibiting most of human cognitive faculties or even superseding them (superintelligence).[35]

---

[30] Wang, 2019, p. 7.

[31] Ibid.

[32] Surden, 2019, p. 1307, referring to P. Norvig & S. Russel, *Artificial Intelligence. A Modern Approach*, Pearson Education, 2003. Interestingly so, Surden continues by claiming "One reason that this characterization of AI is not fully descriptive is that AI has been used to do many activities that humans cannot do. For example, AI technology has been used to spot credit card fraud among billions of transactions using statistical probabilities. […]. If we frame AI as engaging in activities that require human intelligence, we may miss the group of activities that have been automated that humans cannot actually do due to our cognitive limitations".

[33] For an overview of the discussion see, *inter alia*, J. E. Korteling et al., "Human- versus Artificial Intelligence", *Front. Artif. Intell.*, Vol. 4, 2021.

[34] European Council on Foreign Relations, U.E. Franke, Policy Brief "Artificial divide: how Europe and America could clash over AI", 2021, p.5. Available at: https://ecfr.eu/wp-content/uploads/Artificial-divide-How-Europe-and-America-could-clash-over-AI.pdf. Thus, at the same time it is also important to remember that "while AI can already outperform people in spectacular fashion in some domains, like playing board games, in other domains AI is not even competitive with toddlers". Abbott & Sarch, 2019, p. 331.

[35] G. Sartor, "Decisioni algoritmiche tra etica e diritto", in U. Ruffolo (Ed.), *Intelligenza artificiale. Il diritto, i diritti, l'etica*, Giuffré, 2020, p. 66. See also R. Abbott, "Everything is Obvious", UCLA Law Review, Vo. 66, No. 2, 2019, p. 25: "AGI could even be set to the task of self-improvement, resulting in a continuously improving system that surpasses human intelligence—what philosopher Nick Bostrom has termed

One observation is needed at this point: the fact that current AI systems "only" present "weak" intelligence does not entail that an inquiry into responsibility of AI agents is irrelevant. Indeed, as it was efficiently stated, "[p]eople worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world".[36] In fact, narrow AI systems can already operate autonomously (without human supervision) in fulfilling certain purposes. This has been defined as the "control" problem[37] of AI, which can be summarized as follows:

> The rules by which they [*AI Systems*] act are not fixed during the production process, but can be changed during the operation of the machine, *by the machine itself*. This is what we call machine learning. […] Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because *nobody has enough control over the machine's actions to be able to assume the responsibility for them*.[38]

According to Matthias, a *responsibility gap* arises.[39] In particular, the problem raised by Matthias recalls issues which are specific to proving the existence of causality, i.e., the link between the agent's conduct and the harm. Leaving in the background the debate between "techno-pessimists" and "techno-optimists" on whether this *moral* responsibility gap can be

---

Artificial SuperIntelligence (ASI). Such an outcome has been referred to as the intelligence explosion or the technological singularity. ASI could then innovate in all areas of technology, resulting in progress at an incomprehensible rate. As the mathematician Irving John Good wrote in 1965 "the first ultraintelligent machine is the last invention that man need ever make".

[36] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Allen Lane, 2015, p. 286.

[37] Council of Europe, K. Yeung, *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*, DGI(2019)05, 2019, p. 53. Available at: https://rm.coe.int/responsability-and-ai-en/168097d9c5.

[38] A. Matthias, "The Responsibility Gap", *Ethics and Information Technology,* Vol. 6, No, 3, 2019, p.177. Arguing against Matthias' assumption that there is a technology-based responsibility gap, D. W. Tigard, "There Is No Techno-Responsibility Gap", *Philosophy & Technology*, Vol. 1, 2021, pp. 1-19 [emphasis added].

[39] Matthias, 2019, p. 177.

bridged,[40] this research will focus on its the legal conception, namely on whether we are truly facing a "liability gap" and, if so, how it should be addressed.

Returning to the issue of defining AI, the determination of a "legal definition of AI" is currently being addressed by a number of policy makers involved in the booming AI-regulation field. For the purpose of this research, the European Commission High Level Expert Group on Artificial Intelligence's working definition of AI will be adopted:

> Artificial intelligence (AI) systems are *software* (and possibly also *hardware*) systems designed by humans that, given a complex goal, act in the physical or digital dimension by *perceiving their environment* through data acquisition, *interpreting* the collected structured or unstructured data, *reasoning* on the knowledge, or *processing* the information, derived from this data and *deciding the best action(s) to take to achieve the given goal.* AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.
>
> As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems.[41]

To summarize, AI systems are open systems which have the capability of recording large amounts of data, of interpreting them on the basis of specific models, and are able to react and process the data independently, through decision making and problem solver skills. They can therefore act in an autonomous, unpredictable, and not pre-determined way. Artificial intelligences are interactive, adaptable, autonomous, flexible entities.

This research, following the definition and taxonomy proposed by the AI-HILEG, will then use the term "AI Systems" to denote any "AI-based component, software and/or

[40] Tigard, 2021, p. 2.
[41] AI-HLEG, 2018, p. 6.

hardware", whether referring to stand-alone systems or to a single system embedded in a larger one.[42]

1.3 A THREE-RING CIRCUS: E-PERSONHOOD, FAILURES OF CAUSATION,[43] AND GUILTY ROBOT MINDS

One could mention three themes which exemplify the relevance of this investigation: the question of AI systems as legal subjects; the impact of AI's autonomy on the *actus reus* element (specifically, on the causal link between the agent and the harm); and the question of whether AI systems can be said to possess *mens rea*. These themes are referred to respectively as "E-Personhood", "Failures of Causation" and "Guilty Robot Minds" and will be analyzed in this order.

1.3.1 E-Personhood

When discussing AI and criminal liability one must address the issue of conceiving AI systems as *legal* subjects entitled to rights and holders of legal duties. Admittedly, legal subjectivity is a pre-requisite of liability. The debate on the matter has now reached an interdisciplinary dimension and is taking place at the same time in contract, tort, and criminal law. First and foremost, with regards to civil liability, one could refer to the 2017 European Parliament resolution containing recommendations to the Commission on Civil Law Rules on Robotics. The European Parliament explicitly called on the Commission to explore, analyze, and consider the implications of

> *creating a specific legal status for robots in the long run*, so that at least the most sophisticated autonomous robots could be established as having the *status of electronic persons* responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently; [44]

---

[42] Ivi, p. 1.

[43] The expression is coined by Pagallo, 2013, p.73.

[44] European Parliament, *Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103(INL)), 16 February 2017 [emphasis added].

Interestingly so, the European Parliament changed position in a subsequent resolution of 2020, where it stated:

> 7. Notes that all physical or virtual activities, devices or processes that are driven by AI-systems may technically be the direct or indirect cause of harm or damage, yet are nearly always the result of someone building, deploying or interfering with the systems; notes in this respect that it *is not necessary to give legal personality to AI-systems.*[45]

The debate on this matter is currently divided between those pertaining to the "Front of Robotic Liberation"[46], moderates, and skeptics. The distinction will be explained at Chapter 3. For now, it must be underlined that most authors involved in the debate tend to not address explicitly what they mean by "AI". Such an attitude blurs extensively the discussion on AI's personality in criminal law and highlights the importance for this thesis to adopt a definition upfront.

To continue, the most relevant aspect of this debate revolves around two arguments: the "AIs Are Not Human" and the "Missing-Something" arguments.[47] According to the first argument, modern legal systems are rooted in an anthropocentric standpoint, stemming from the teachings of the Enlightenment, and therefore legal personhood of AI agents should be refuted. This argument could be summarized with the following maxim: "*Das Strafrecht ist von Menschen für Menschen erdacht worden*".[48]

Thus, legal systems have historically addressed issues regarding the attribution of responsibility arising from "non-human" behavior: think of wrongdoing committed by animals or by subjects which lack the legal capacity to behave criminally (for a time, slaves, but also children and insane offenders). In point of fact, in certain legal systems being human

---

[45] European Parliament, *Resolution of 20 October 2020*.

[46] Pagallo, 2013, p. 155. The issue will be analyzed in depth in Ch. 3.

[47] L. B. Solum, "Legal Personhood for Artificial Intelligences", *N.C. L. Rev.*, Vol. 7, 1992, pp. 1258-1276.

[48] L. Quarck, "Zur Strafbarkeit von e-Personen", *Zeitschrift für Internationale Strafrechtsdogmatik*, Vol. 2, 2020, p. 67. As stated also by Seher: "*Das Strafrecht ist dazu gemacht, Menschen dafür zu sanktionieren, dass sie grundlegende Regeln des rechtlichen Zusammenlebens verletzt haben […] Die Adressaten des Strafrechts sind Menschen als Teilnehmer, Mitwirkende und Unterworfene des Normensystems „Recht" – Personen im Recht*", G. Seher, "Intelligente Agenten als 'Personen' im Strafrecht?", in S. Gless & K. Seelmann (Eds.), *Intelligente Agenten und das Recht*, Vol. 9, 2016, Nomos, pp. 45-46.

is neither a *necessary* nor a *sufficient* condition for being qualified as a legal subject in the field of criminal law, according to the *societas delinquere potest* principle.[49]

In view of that, as stated by Chopra and White:

> Arguments for advancing personhood for artificial agents need not show how they may function as persons in all the ways that persons may be understood by a legal system, but rather that they may be understood as persons *for a particular purpose* or set of legal transactions. For the law does not always characterize entities in a particular way for all legal purposes.[50]

Could AI Systems, then, be regarded as *Rechtspersonen* for the purposes of criminal law? For example, an important difference between AI systems and a legal person is that the latter defines itself as such (and acquires personhood upon constitution). The same cannot be said with regards to AI systems.

Certainly, "[n]ew technologies result in the *implosion of the legal subject as we knew it* so that the legal subject and the human, natural person, conceptually different though they are, no longer necessarily coincide".[51] According to Hildebrandt, "there is no categorical legal answer to the question whether an autonomous computational system […] should be given legal personhood. That question is a *political question* that must be answered by a legislature weighing the advantages and disadvantages of such a move. […]".[52]

Moving forward, the second argument, i.e., the "Missing-Something" Argument, is of utmost interest with regards to this research's field of investigation. It could be summarized as follows: currently AI systems do not possess consciousness, intentionality, and morality, hence they lack the necessary preconditions to be treated as legal subjects and for attribution

---

[49] Corporate criminal liability currently exists in a number of jurisdictions such as the United States, England, Australia, Canada, Finland, Denmark, France and the Netherlands. For an overview of modern systems of criminal justice and the different models of assessing responsibility for crimes committed by corporations, see: C. De Maglie, "Models of Corporate Criminal Liability in Comparative Law", *Wash. U. Global Stud. L. Rev.*, Vol. 4, 2005.

[50] S. Chopra & L. White, *A legal theory for autonomous artificial agents*, University of Michigan Press, 2011, p. 156.

[51] J. Gaakeer, " 'Sua cuique persona?' A Note on the Fiction of Legal Personhood and a Reflection on Interdisciplinary Consequences", *Law & Literature*, Vol. 28, No. 3, 2016, p. 303.

[52] M. Hildebrandt, *Law for Computer Scientists and Other Folk*, Oxford University Press, 2020, p. 246.

of criminal liability[53]. Leaving the issue of morality and consciousness in the background for now, let us focus on three elements: intent, free will, and autonomy.[54]

Being *sui juris* in criminal law implies the notion that a "subject of the law must understand the nature of the act it commits".[55] Consequently, the "most damning"[56] objection to an artificial agent possessing free will is that it is "just a programmed machine".[57] Thus, one should not take comfort in the assertion that humans are not programmed, while instead artificial agents unequivocally are.[58] As a matter of fact, according to the most recent neuroscientific studies, our decisions are nothing but the result of *encoded* brain activity of the prefrontal and parietal cortex.[59] As stated effectively by Gless and Weigend, robots might soon appear to us as no more "remotely controlled" than humans, who take their (supposedly) free decisions within an inscrutable web of influences that come from their genetics, their education, and the social environment. Yet, we still define them as "freely" responsible beings.[60]

Finally, it is argued that the very existence of AI systems raises the question of what exactly is human about humans ("*was genau das Menschliche am Menschen ist*").[61] This thought will acompany us till the very end of this research.

1.3.2 Failures of Causation

The issue of whether an AI criminal conduct fulfills the *actus reus* element of an offense is sometimes discarded easily in relevant literature. The question of the capacity of intelligent agents to act under criminal law is also a matter of definition: in a causalistic-external view, which defines every arbitrary bodily movement as an action, AI-systems can be regarded as

---

[53] Pagallo, 2013, p.157.
[54] Chopra & White, 2011, p.173.
[55] Ivi, p.165.
[56] Chopra & White, 2011, p.165.
[57] Ibid.
[58] Ibid.
[59] Ibid.
[60] "*Möglicherweise erscheinen uns Roboter ja bald nicht mehr stärker „ferngesteuert" als Menschen, die ihre vermeintlich freien Entscheidungen in einem undurchschaubaren Geflecht von Einflüssen aus Genetik, Erziehung und sozialer Umwelt treffen und die wir dennoch als „frei" verantwortlich definieren*". S. Gless & T. Weigend, "Intelligente Agenten und das Strafrecht", *ZSTW*, Vol. 126, No. 3, 2014, p. 568.
[61] Ivi, p. 588.

agents. The more "substantially charged"[62] the concept of act is, the more self-consciousness is read into it, the less intelligent agents will meet the requirements set by criminal law.[63] Notably, the concept of *actus reus* generally includes the issues of conduct and causation. It has a heterogeneous nature. As a matter of fact, the term can be used to refer to a "collection of entirely distinct doctrines with different functions".[64] Far from turning a blind eye from those who criticize that an AI can even "act" in a *Tatbestand* perspective,[65] this paragraph will now focus on the so-called "failure*[s]* of causation". [66]

Pagallo applies the expression "failure of causation" to describe the disruptive effect of the autonomy of AI agents on the link between agency and the occurred negative outcome (in a cause and effect analysis: if A, then B).[67] He states laconically, but accurately: "Matters of legal causation, are, traditionally, a nightmare for legal scholars" .[68]

The problem of ascertaining causality between the AI system and the harmful event arises because this kind of explanation is linked to the possibility for man to fully dominate (etiologically) a certain event, something that in certain AI applications may not be possible. It can be demonstrated that a system starting from a set of inputs has produced some outputs, but it might not be possible to explain why and how. Sometimes, we might not even be able to understand which inputs had a role in obtaining the output. Indeed, as it was already highlighted before, in cases of AI-based crime "the distance between a human action and its consequences increase exponentially".[69]

---

[62] Gless & Weigend, 2014, p. 572.

[63] "*Letztlich dürfte die Frage nach der (strafrechtlichen) Handlungsfähigkeit Intelligenter Agenten eine Frage der Definition sein: Bei „kausalistischer", bloß äußerlicher Betrachtung, die jede „willkürliche Körperbewegung" als Handlung definiert, sind sie durchaus als Handelnde anzusehen. Je stärker man den Begriff der Handlung substantiell auflädt, je mehr an selbstbewusster Zielbestimmung man in ihn hineinliest, desto weniger können Intelligente Agenten den Voraussetzungen der Handlungsfähigkeit genügen*". Ibid.

[64] J. Keiler & D. Roef, "Principles of Criminalisation and the Limits of Criminal Law", in J. Keiler & D. Roef (Eds.), *Comparative Concepts of Criminal Law*, 3rd Ed., Intersentia, 2019, p. 61.

[65] D. Lima, "Could AI Agents Be Held Criminally Liable? Artificial Intelligence and the Challenges for Criminal Law", *South Carolina Law Review*, Vol. 69, Issue 3, 2018, p. 677. This issue will be dealt with into detail in Ch. 6.3.

[66] Pagallo, 2013, p.73.

[67] Pagallo, 2013, p.73.

[68] Ibid.

[69] Hildebrandt, 2014, p. 190.

This study identified three factors which bring about failures of causation: (1) the problem of many hands; (2) the black box problem and (3) shortcuts.[70] The four factors will be briefly touched upon in this order, leaving the more in-depth analysis for later on in this research.

(1) The Many Hands Problem

Borrowing an expression which was coined in the field of philosophy and moral responsibility,[71] the expression "problem of many hands" refers to the fact that the development of AI systems is often the result of the combinations of actions by numerous individuals and the outcome of a long, and complex, chain of individual efforts.[72] Consequently, when harm occurs, it is very hard – if not impossible – to identify the individual cause, and therefore the liable individual, behind the event. The problem of many hands is particularly evident in cases of open-source software. It poses significant challenges to the ascription of criminal liability which is, by nature, based on *individual*, and not collective, responsibility.

(2) The Black Box Problem

The "Black Box" problem regards specific AI-systems (based, for example, on Deep Neural Networks, DNNs) which solve problems in an *opaque* way.[73] It can be defined as "an inability

[70] Council of Europe & Yeung, 2019, p. 12.

[71] For an in-depth philosophical analysis of the "problem of many hands", see I. van de Poel, L. Royakkers, & S. D. Zwart, *Moral Responsibility and the Problem of Many Hands*, Routledge, 2018.

[72] "First identified in the context of information technology by philosopher of technology, Helen Nissenbaum, the problem of 'many hands' is not unique to computers, digital technology, algorithms or machine learning. Rather, it refers to the fact that a complex array of individuals, organisations, components and processes are involved in the development, deployment and implementation of complex systems, so that when these systems malfunction or otherwise cause harm, it becomes very difficult to identify who is to blame, because such concepts are conventionally understood in terms of individualistic conceptions of responsibility. In other words, causal responsibility is necessarily distributed where complex technological systems are concerned, diluting causation to mere influence". Council of Europe & Yeung, 2019, p. 63.

[73] For a clean cut explanation of the black box problem, see D. Castelvecchi, "Can we open the black box of AI?", *Nature*, Vol. 538, 5 October 2016. Available at: https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731. For a thorough overview of the impact of black-box systems on intent and causation, see: Y. Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation", *Harvard Journal of Law & Technology*, Vol. 31, 2018.

to fully understand an AI's decision-making process and the inability to predict the AI's decisions or outputs".[74] It entails that it might be impossible for the designer or the user of the program to foresee AI-behavior after the system is deployed, and to reconstruct why it acted in a specific way *ex post*.

DNNs are built on layers of interconnected artificial neurons which work together. A DNN will make decisions using an intuitive decision-making process and it will learn through experience. If we think of an analogy, it is the same as when you "know" how to ride a bike because you learnt by attempting, and falling, over and over again.[75] It is impossible for someone to explain to external subjects (for example by describing the correct movements for maintaining balance while on a bike) how such knowledge was reached. One could say that"[y]ou use your brain all the time; you trust your brain all the time; and you have no idea how your brain works".[76]

(3) Shortcuts

Shortcut is a phenomenon that occurs in ML, specifically in self-supervised ML (a model which learns by itself: it is fed raw data without any label description from humans, and it is then told, for example, to classify this data). Sometimes, the model relies on a simple characteristic in a dataset (for example, the green grass) rather than learning the true meaning of that data. This phenomenon is referred to as a "shortcut" and is very difficult to detect, since the outputs of the AI systems are, at a first glance, correct. The result of shortcuts is inaccurate predictions. We might question how a judge shall apply the standard rule on attribution, that is, the operation through which she selects the legally relevant cause of a harmful event, in cases where the AI system takes a shortcut. What if it is not possible to reconstruct the web of causation *ex-post*? Which scientific theory should the judge rely on in order to ascertain liability?

Unquestionably, the three factors mentioned above emphasize the disruptive character of failures of causation.

---

[74] Bathaee, 2018, p. 905.
[75] Ivi, p. 902.
[76] P. Baldi, quoted in Castelvecchi, 2016, p. 23.

1.3.3 Guilty Robot Minds

Some of the most fundamental problems in the attribution of criminal liability to AI systems arise with regards to the subjective element of offenses, also referred to as the mental state (*mens rea*) element. This represents – at the state of the art – one of the most heated debates in the (newborn) field of AI criminal theory.

Indeed, if one assumes that intelligent agents can fulfill the *actus reus* of a crime, the next natural question to be asked is whether intelligent agents can be *guilty*.[77] In this regard, scholars mainly base their opinions drawing on comparisons with corporate liability.[78] Actually, according to American criminal legal doctrine, mental states can be imputed to corporations following the *respondeat superior* theory and, to a certain extent, such reasoning could be applied to ascribe culpable mental states of "humans-in-the-loop" (AI developers, owners, users) to the artificial agent. As argued by Hallevy,

> […] there is no substantive legal difference between the idea of criminal liability imposed on corporations and on AI entities. It would be outrageous not to subordinate them to human laws, as corporations have been. Models of criminal liability exist as general paths to impose punishment. *What else is needed*?[79]

On the contrary, others argue that "[u]nlike a corporation, which is literally composed of the humans acting on its behalf, an AI is not guaranteed to come with a *ready supply of identifiable human actors* whose mental states can be imputed",[80] and suggest the development of a set of specific AI-strict-liability offenses, therefore allowing for punishment of AI without requiring mental states.

Leaving the question of attributing *mens rea* directly to AI-systems for later elucidation, there is another battlefield where the academic discussion has gained traction: negligence crimes for humans-behind-the-machine. Should legal systems create new negligence crimes for developers/producers/users of AI systems? Is it feasible to speak of a sort of *culpa in*

---

[77] Gless & Weigend, 2014, p. 573.

[78] The matter is addressed in depth at Ch. 6.4.

[79] G. Hallevy, "The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control", *Akron Intellectual Property Journal*, Vol. 4, Issue 2, 2010, p. 201[emphasis added].

[80] Abbott & Sarch, 2019, p. 352 [emphasis added].

programmando as a twin sister of the already-existing *culpa in vigilando* doctrine? Or should we look elsewhere, namely tort law?

1.4 RESEARCH QUESTION & STRUCTURE OF CHAPTERS

The research will be guided by the following main research question: *to what extent is a theoretical framework of criminal law for liability of non-human agents needed and feasible?*

First and foremost, the following introductory chapter will be concluded with a list of real-life or hypothetical scenarios of "AI going wrong",[81] which will be recalled during the research. The examples chosen are purposively trivial and are meant to introduce the issues discussed in the thesis in an unpretentious, yet straightforward, manner.

AI is a multifaceted concept which is rooted in a different field than legal doctrine, i.e., computer science. Finding a univocal definition is challenging, even in its home-field. For this reason, this study starts by addressing such problems in Chapter 2. It targets the following sub-questions: *why is the issue of defining AI an issue? What are the basic functions of AI?* After providing a short history of AI and of its most modern applications, Chapter 2 gives an overview of the different definitions put forth since the late 1950s and of its basic functioning. It will then be concluded with the adoption of a working definition.

Chapter 3 accounts for the ongoing scholarly discussion on AI and criminal law. The analysis is necessary in order to position this thesis in the relevant debate. It addresses the following sub-questions: *What is the state of the art of the scholarly debate on AI and criminal law? Which are the most recurrent questions and what are the answers? Which aspects are being neglected?*

Chapters 4, 5, and 6 focus on *ascription* and address the following sub-question: *can liability for AI Crimes be attributed to the AI agent itself?* To do so, Chapter 4 introduces different concepts, ranging from the "Paperclip Maximizer" thought experiment and the notion of utility function, to the similarities between AI incidents and other technological failures, such as aviation incidents (and their criminalization). Moreover, it focuses on different theories of criminalization, i.e., the framework in which happens the selection of criteria for the ascription of an offense.

---

[81] For an updated database of AI related incidents, see the AI Incident Database, available at: https://incidentdatabase.ai/.

Chapters 5 and 6 represent the beating heart of this thesis. The inquiry in Chapters 5 and 6 is focused on understanding if modern models of criminal liability should be, and can be, adapted to algorithmic harm in a way that is compatible with principles related to criminal punishment, such as the principle of legality and the principle of blameworthiness.

Specifically, Chapter 5 focuses on criminal capacity, i.e., the set of conditions required to be an addressee of criminal law. First, it discusses the contours and relevance of criminal capacity in criminal law. Then, it addresses whether AI systems can have personhood in criminal law. Chapter 6 analyzes AI crime both in a *mens rea* and *actus reus* direction. With regards to *mens rea,* Chapter 6.2. inquiries into whether *mens rea* can be imputed directly on the AI. Moreover, it analyzes the liability of the "humans-behind-the-machine". When it comes to the specific discussion of *human* liability, the Chapter addresses mainly issues connected to *negligent* offenses, as it is deemed the most problematic, hence worthy of investigation, field. With regards to *actus reus* (Chapter 6.3.), attention is given to whether an AI system can act in a criminally relevant way and to the impact of "failures of causality" on the reconstruction of the causal link. Chapter 6.4. touches upon models of corporate liability. This analysis is mandated by the fact that a very conspicuous number of scholars draws upon the analogy with corporations in order to base their arguments on criminal liability for AI systems. Thus, this research considers the analogy with a critical eye, assessing its strengths and flaws. After doing so, the research briefly stretches its frontiers to discussing Diamantis' "algorithmic corporate misconduct"[82] and "corporate algorithmic harm"[83] theories.

Chapter 7 presents an overview of adopted (or proposed regulations) on criminal liability of AI updated to February 2023. It addresses the following sub-question: *what is the state of the art of criminal law regulation on AI*? The latitude of the analysis in the Chapter is global. Specifically, it tackles: A) The CoE's CDPC and the drafting of an "Instrument on Artificial Intelligence and Criminal Law";[84] B) the Singapore Penal Code Review Committee Report of 2018[85] and the Report on "Criminal Liability, Robotics and AI systems" drafted

[82] M. E. Diamantis, "The Extended Corporate Mind: When Corporations Use AI to Break the Law", *N.C.L. Rev.*, Vol. 97, 2020.
[83] Ibid.
[84] As it will be discussed in Ch. 7.2., the CDPC has not adopted said instrument yet.
[85] Singapore Penal Code Review Committee, Report, 2018 ("PCRC Report"). Available at: https://www.reach.gov.sg/-/media/reach/old-reach/2018/public-consult/mha/annex--pcrc-report.ashx.

by the Singapore Law Commissions of 2021;[86] C) the legislative reform of the French Road Act;[87] D) the "Automated Vehicles: joint report" drafted by the Law Commission of England and Wales and by the Scottish Law Commission;[88] E) the amendment of the German Road Traffic Act.[89]

Finally, Chapter 8 presents the conclusions of the thesis and delineates directions of future inquiry. To do so, it will retrace the sub-questions of the chapters.[90] Consequently, it will first return to the issue of defining AI and its repercussions on criminal law. Then, it will situate this research in the realm of the scholarly literature on AI and criminal law analyzed in Chapter 3. Next, it will deliberate on the attribution of criminal liability to AI systems, by focusing specifically on why humans tend to hold AI systems to higher moral standards; on theories of retributivism and deterrence; and on the analogy between AI liability and corporate criminal liability. Moreover, the possibility of liability of humans *for* machines will also be reviewed. Finally, Chapter 8 will return to the main RQ, outline perspective for future avenues of research, and conclude with the challenges which were encountered in the study.

1.5 METHODOLOGY

As already outlined, in order to embark in any kind of theoretical and critical dialogue on this topic it is necessary to first review the authority and status of the legal doctrine on the issue. Hence, the research provides an extensive literature review at Chapter 3, which focuses on

---

[86] Singapore Academy of Law, Law Reform Committee, "Report on Criminal Liability, Robotics and AI Systems", 2021 ("SAL Report"). Available at:

[87] Ordonnance n° 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation (TRAT2034523R, JORF n°0089 du 15 avril 2021, Texte n° 36), 2021.

[88] Law Commission of England and Wales Law Commission No. 404, Scottish Law Commission Scottish Law Commission No. 258, "Automated Vehicles: Joint report", HC 1068 SG/2022/15, 25 January 2022 ("Law Commissions Report"). Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2022/01/Automated-vehicles-joint-report-cvr-03-02-22.pdf.

[89] Bundestag, Gesetz zur Änderung des Straßenverkehrsgesetzes und des Pflichtversicherungsgesetzes – Gesetz zum autonomen Fahren, Bundesgesetzblatt Jahrgang 2021, Teil I Nr. 48, 3108.

[90] Chapter 2 – *why is the issue of defining AI an issue? What are the basic functions of AI?*; Chapter 3 – *What is the state of the art of the scholarly debate on AI and criminal law? Which are the most recurrent questions and what are the answers? Which aspects are being neglected?*; Chapter 4 – *What does ascribing criminal liability to an AI system entail?*; Chapter 5 – *Can an AI system be treated as a criminal agent?*; Chapter 6 – *Can an AI system fulfil the* mens rea *and* actus reus *requirements of a criminal offense?*; Chapter 7 – *what is the state of the art on criminal law regulation on AI?*.

the current debate on the issue through literature and hence clarifies the scope of the research that has been conducted in this area.[91]

This literature review has a double soul. First, it represents the theoretical background of the thesis. In this sense, it is essential for various reasons such as: to determine what has been written on the topic of the thesis; to identify trends or patterns in that research area; and to identify questions that require further inquiry.[92] Second, it represents an "original and valuable work of research in and of itself",[93] as it created "a solid starting point for all other members of the academic community that are interested"[94] to this specific overarching topic.

Chapter 3 includes doctrinal studies from diverse countries. The reasons for this large selections are manifold. Firstly, the scope of the debate is international, so disregarding pieces of relevant literature would diminish the research's value; secondly, the question of AI criminal liability is being addressed through a shared language, that is, through doctrine on the common substantive aspects of criminal law. The literature review is indeed crucial for the purpose of building a valid foundation with regards to the existence and relevance of the problems discussed in the thesis.

The search strategy which was adopted for the literature review is the following. Relevant databases[95] were searched using words such as criminal liability, artificial intelligence, *mens rea*, *actus reus*, criminal legal framework, and accountability as search terms in the three languages mentioned above. The results of the analysis are based on approximately 100 texts in three languages (Italian, English, and German) published from

---

[91] The nature of a literature review has been described as follows: "The literature review is the part of the thesis where there is an extensive reference to related research and theory in your field. It is where connections are made between the source texts that you draw on and when you position yourself and your research among their sources. It is your opportunity to engage in a written dialogue with researchers in your area while at the same time showing that you have engaged with, understood and responded to the relevant body of knowledge underpinning your research. The literature review is where you identify the theories and previous research which have influenced your choice of research topic and the methodology you are choosing to adopt. You can use the literature to support your identification of a problem to research and to illustrate that there is a gap in previous research, which needs to be filled. The literature review, therefore, serves as the driving force and jumping-off point for your research investigation". G. Pare et al., "Synthesizing information systems knowledge: A typology of literature reviews", *Information & Management*, No. 52, 2015, p. 183.

[92] Ibid.

[93] Pare, 2015, p. 183.

[94] Ibid.

[95] Including, but not limited to De Jure, HeinOnline, SSRN, ArXiv, Research Gate, Academia.

when the topic gained popularity (circa 2010) until today. The analyzed texts were then categorized into three groups, which reflect the sides of the debate: skeptics, moderates, and expansionists.

The main research question is addressed in Chapters 4 to 6 by adopting a doctrinal methodology. When discussing the dogmas of traditional criminal law, the research will, at times, refer to different legal systems (mostly Italy, Germany, and the US). These countries are of interest for two main reasons: first, it was presumed that they would represent the country of provenance (or of reference) of many of the authors analyzed in Chapter 3; second, they are prototypes of different approaches to issues of criminal liability, especially with regards to corporate criminal liability.[96] Yet, the analysis in Chapters 4, 5, and 6 is *not* based on a comparative approach. Such method was discarded since – upon an initial survey – it became evident that there was little to nothing debate on the object of investigation at a *legislative* level. Only the existence of such a debate would have rendered the comparative research a reasonable exercise.

The initial assumption on the (still) little relevance of hard law regulation when it comes to AI and criminal liability is confirmed in Chapter 7. This Chapter has a wider breath, since it does not follow a geographical criteria in the selection of the samples. Rather, it adopts a subjective one: it includes any type of legislation, policy, or report which addresses the general field of AI and criminal liability. The 5 samples examined therein are then compared with each other, and assessed according to their similarities, differences, and relevance.

[96] Germany at the moment provides only for administrative corporate criminal liability – *Ordnungswidrigkeitengesetz* – even though there are recurrent discussion on the introduction of corporate criminal liability via new legislation; Italy has adopted a hybrid system with Law 231/2001; in the United States corporations can be held directly criminally liable.

1.6 AIs GOING BAD – EXAMPLES

Example A – The Driverless Car

An autonomous vehicle produced by the company Trust'n'ride, carrying a safety driver, runs over a woman who is walking a bicycle while crossing the street.[97] The woman was jay walking and the road was dry and illuminated by street lighting. The system had failed to identify the pedestrian until 5.6 seconds before the impact, as it was not trained for detecting jaywalking pedestrians. The driver operator was looking at her phone and started steering the wheel to avoid the collision 0.2 seconds before hitting the pedestrian. According to the findings of an independent governmental authority, the probable cause of the crash was the vehicle operator's failure to monitor the driving environment and the operation of the automated driving system because she was distracted by her personal cell phone. Moreover, Trust'n'ride's inadequate safety risk assessment procedures and ineffective oversight of vehicle operators contributed to the crash. Interestingly so, the system design had precluded the activation of the emergency braking system for collision mitigation (it relied instead in the human's intervention to avoid a collision). The County Attorney of Rose Hills decided not to press criminal charges against the company and instead pursued charges against the backup driver who is eventually indicted with one count of negligent vehicular homicide. Should the backup driver be convicted?

Example B – Predicting Suicide

An AI system is used for predicting suicide attempts by applying machine learning to electronic health records of patients seen for any reason.[98] Imagine the system, which has proven to have an accuracy of 80%, predicts that a patient admitted to the emergency room for a sprained ankle is at high risk of committing suicide. Said evaluation is done based on hospital-admission data such as age, gender, past diagnosis, socioeconomic status, and

---

[97] The example is inspired by the killing of Elaine Herzberg in Tempe, Arizona. The Guardian, "Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian", 19 March 2018. Available at: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

[98] Several studies with the purpose of developing accurate suicide attempt risk models have already been developed and currently being used in a number of fields, see C. G. Walsh, J. D. Ribeiro & J. C. Franklin, "Predicting Risk of Suicide Attempts Over Time Through Machine Learning", *Clinical Psychological Science*, Vol.5, Issue 3, 2017; C. G. Walsh et al., "Prospective Validation of an Electronic Health Record–Based, Real-Time Suicide Risk Model", *JAMA Netw Open*, Vol. 4, No. 3, 2021. Other models include: REACH VET (Recovery Engagement and Coordination for Health—Veterans Enhanced Treatment) from the Veterans Health Administration and Army STARRS (Study to Assess Risk and Resilience in Service members).

employment status, to mention a few.[99] Some of these factors are not usually considered by doctors in evaluating suicide risks and most of the times such evaluation is not even conducted on patients admitted for this type of injury. What if the algorithm foresees a high suicidal risk and the doctor's clinical opinion differs, leading to a discharge of the patient and his suicide? Should the doctor be liable for the death of her patient? If we consider the opposite scenario, what if the algorithm does not predict a suicide? Who is to blame?

Example C – Evil AI

Meet Norman, the world first psychopath AI.[100] In 2018, researchers at MIT created an AI system capable of perform image captioning (a deep learning method developed to generate a textual description of an image). The algorithm was trained on image captions taken from a dark thread on the Subreddit platform which is dedicated to document and observe death. Then, the system was subjected to the Rorschach inkblots test. This test is normally used to analyze personality traits and to diagnose certain mental illnesses. The results of Norman were then compared with the responses of a standard image captioning neural network trained on a MSCOCO (Microsoft Common Objects in Context) dataset.[101] The regular AI saw a close up of a vase with flowers, Norman saw a man shot dead. The aim of the experiment was to prove that there is no such thing as "inherently bad" AI systems, just systems trained with bad data. According to most, this experiment is proof of the dangers of AI trained on biased data falling into the "wrong hands". If a system designed as Norman went rogue and committed a crime, could it be said that it possessed *mens rea*? Can AI plead insanity?

Example D – Moral Dilemmas

1 – The Trolley Problem. An automated vehicle (AV) driving in autonomous mode, carrying a human, is driving at high velocity on a road where two seriously injured people (A and B) are lying. If the AV skews right to avoid running over (and killing) A and B, it will hit C and D who are standing at the side of the road. Based on (very fast-paced) calculations, including the almost certain probability of injuring C and D, and the likely risk of killing A

[99] See O. Goldhill, "Machines know when someone's about to attempt suicide. How should we use that information?", *Quartz*, 5 September 2018. Available at: https://qz.com/1367197/machines-know-when-someones-about-to-attempt-suicide-how-should-we-use-that-information/.

[100] Available at http://norman-ai.mit.edu/.

[101] Dataset containing 328k photos of 91 object's types that would be easily recognizable by a 4 year old. See T. Lin et al., "Microsoft COCO: Common Objects in Context, Computer Vision – ECCV", *Lecture Notes in Computer Science*, Springer, Vol. 8693, 2014.

and B, it decides not to swerve. As a consequence, A and B die from the impact. Could it be argued that the AV's act was justified or excused?[102]

2 – The Burning Room Dilemma. Imagine that an object of value is trapped in a room that is on fire and that a human, who does not want to put herself in danger by retrieving the object, instructs a capable robotic companion to get the object from the room ('*short_Grab'* action) and bring it to safety, thus risking self-destruction. The agent could also take a longer route (*'long_Grab'* action), which avoids the fire, but that entails a 0.0.5 probability that the object may be destroyed during the time it takes for the robot to complete the route. It is assumed that the robot is unsure of whether the human values the object more than it values its own safety. What decision should the robot-agent make in this critical scenario?[103]

3 – The drowning child.[104] Let us imagine that a delivery robot is walking on the beachside, when it hears a child crying for help because she is drowning. Suppose that the robot will make a series of calculations which include the option of stopping to pull the child out of the water (which leads to a 10% change that it might fall inside and destroy itself) and the option of not doing anything (which leads to a 90% chance that the child will die). The robot knows with a 95% probability that the drowning child is actually a human child who is in imminent danger. Let us assume that the robot decides not to take any action because it deems its task of delivering a package more important and, as a consequence of its omission to act, the child drowns. Should the robot be liable?
Let us assume now that the robot did not react to the child's cry for help because it is convinced that it is already in the process of saving another drowning child, who instead is already dead. Should a change in the robot's motive to act (i.e., from deciding that delivering a package is more important than saving the drowning's child life, to mistakenly decide to continue saving another's child life) affect evaluations regarding its liability?

---

[102] The example is adapted from E. Hilgendorf, "The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems", in E. Hilgendorf and J. Feldle (Eds), *Digitization and the Law*, Robotik und Recht, Vol. 15, Nomos, 2018, p. 75.
[103] The example is taken from D. Abel, J. MacGlashan & M.L. Littman, "Reinforcement Learning as a Framework for Ethical Decision Making", *AAAI Workshop: AI, Ethics, and Society*, 2016.
[104] The example is adapted from Y. Hu, "Robot criminals", *U. Mich. J.L. Reform*, Vol. 52, 2019, p. 500.

2 DEFINING ARTIFICIAL INTELLIGENCE

Do AIs dream of electric sheep?
Philip K. Dick (1968)

2.1 Introduction: to Define or not to Define? – **2.2** A Short History of AI – **2.3** Definitions of AI: an Overview – **2.4** Gaining a Basic Understanding of Modern AI Systems – **2.5** Adopted Working Definition – **2.6** Conclusion

2.1 INTRODUCTION: TO DEFINE OR NOT TO DEFINE?

"AI is learning how to create itself',[105] recited sensationally the MIT Technology Review – one of the most renowned online newspapers on technology – on May 27, 2021. The article was commenting a research named POET (acronym for Paired Open-Ended Trailblazer) that is currently being conducted in the Uber AI Labs by Ruy Wang.[106] POET is an open-ended algorithm, meaning that it works and progresses without pre-established goals.[107] The system is being tested in a 2-D bipedal-walking-obstacle-course where the algorithm

[105] W.D. Heaven, "AI is learning how to create itself. Humans have struggled to make truly intelligent machines. Maybe we need to let them get on with it themselves", *MIT Technology Review*, 27 May 2021. Available at: https://www.technologyreview.com/2021/05/27/1025453/artificial-intelligence-learning-create-itself-agi/.

[106] R. Wang et al., "Paired Open-Ended Trailblazer (POET): EndlesslyGenerating Increasingly Complex and Diverse Learning Environments and Their Solutions", *ArXiv abs/1901.01753*, 2019.

[107] The definition of open-endness is debated amongst AI researchers. Evolutionary algorithms design represents one of the growing fields of study in computer science now. For the purposes of understanding the functioning of POET, it is relevant to state that "[i]n an open-ended learning process, an agent or robot must solve an unbounded sequence of tasks that are not known in advance", hence without knowing the given domain (comprising of its states, actions and rewards). S. Doncieuxet et al., "Open-Ended Learning: A Conceptual Framework Based on Representational Redescription", in *Frontiers in neurorobotics*, Vol. 12, p. 59, 2018. Available at: https://doi.org/10.3389/fnbot.2018.00059. For an in-depth analysis of concept of open-ended evolution as applied to AI, see O. Stanley Kenneth; "Why Open-Endedness Matters", *Artif Life*, Vol. 25, No. 3 2019, pp. 232–235; N. Packard et al., "An Overview of Open-Ended Evolution: Editorial Introduction to the Open-Ended Evolution II Special Issue", *Artificial life*, Vol 25, No. 2, 2019, pp. 93–103.

simultaneously creates problems (the paths and the obstacles, e.g., a flat terrain or a hill), solves them (e.g., by making the two-legged figure jump or run), and learns from them (Fig. 1). "Given any problem space with the potential for diverse variations, POET can blaze a trail we through it", [108] says Wang. Thus, if look at POET's presentation video, we are not struck by the bot's complexity or problem-solving capabilities.[109]

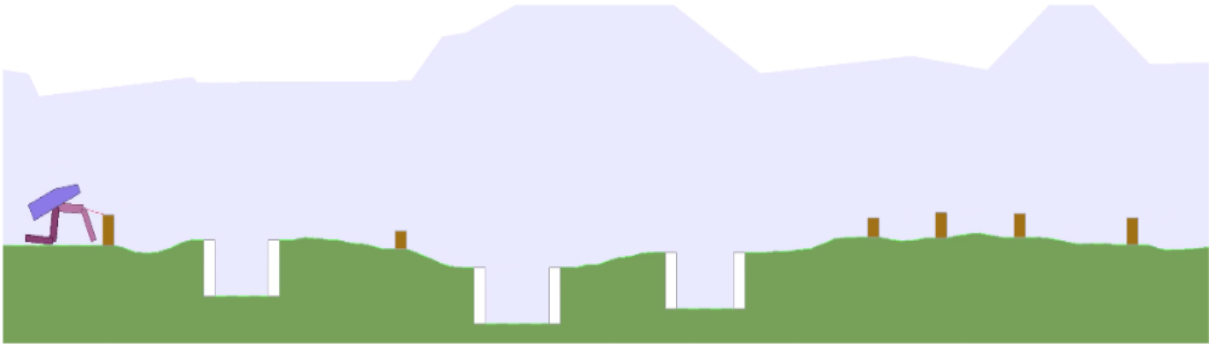

Figure 1 An example of POET'S bipedal-walking obstacle course. Source: R. Wang et al., "Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions", arXiv:1901.01753v3, 2019.

In fact, what one notices *prima facie* is just a tiny limping stick figure with a wedge-shaped head that attempts to walk and to overtake obstacles, dragging its "knee" on the ground and stumbling from time to time (Fig.1).[110] In all fairness, nothing as razzle-dazzle as hinted by the article's presentation. Indeed, the AI system we are able to perceive in this image seems far from carrying out complex tasks, such as screening millions of genetic compounds to develop new drugs[111] or operating on a patient with extreme precision in a short matter of time through a robotic arm.[112]

---

[108] Wang, 2019.

[109] Ibid.

[110] Heaven, 2021.

[111] Think for example of AtomNet, a drug discovery platform developed by Atomwise, which by applying deep convolutional neural networks is able to predict the bioactivity of small molecules for drug discovery application. Notably, thanks to one of its ongoing projects it found a drug candidate that might be applied to combat Ebola and multiple sclerosis. For a deeper analysis, see: I Wallach, M. Dzamba & A. Heifets., "AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery", *ArXiv abs/1510.02855*, 2015.

[112] For example think of the system developed by Asensus Surgical which performs a type of telesurgery named Performance-Guided Surgery: through augmented intelligence a robotic-assisted platform is able to perceive (computer vision), learn (machine learning) and assist (clinical intelligence) in minimally invasive surgeries such as laparoscopy. Such system could allow a surgeon sitting in a room to operate in more than 2 or 3 operating rooms at the same time, be it in the same hospital or in another country. See

Yet, if one takes a closer and more informed look, it seems as POET poses even more issues than other AI systems. It is an instance of using AI to make AI and, as such, it has the revolutionary potential of changing the narrative: it entails using machine learning (ML) algorithms not only as tools to solve foreseen problems, but also to generate their own problems and solutions: "[u]nlike algorithms in, for example, machine learning that learn to solve problems we *ask them to solve*, open-ended algorithms could produce surprises beyond our imagination *without the need to ask*".[113] Could we substitute the word "surprises" with "risks"? Or even "harm"? As some have said, this might be first step towards achieving Artificial General Intelligence, also referred to as 'super intelligent machines'.[114]

The examples above helped to point our attention to one of the existential issues when studying AI: what is AI exactly? Is it the awkwardly walking two-legged bot or the sophisticated driverless car? As the scope of this thesis is to discuss a liability framework involving AI-actions, what should judges, legal professionals, and regulators consider as AI in a criminal law context? In essence: *why is the issue of defining AI an issue?*[115]

On a preliminary note, it is relevant to stress that avoiding ambiguities in law is crucial: legislators and policy makers strive to create technical language that is void of polysemous words. When doing so, they act as lexicographers and create a technical vocabulary that might be extremely different from the ordinary 'popular' one.[116]

When it comes to AI, the urge for definitions is enhanced by two factors. First, there is no overall accepted definition of AI[117] – even though it is possible to identify systematic efforts in this direction by the first players in the AI-regulation field, such as the European Union.[118] Second, in order to define AI it is necessary to touch upon technical concepts

---

G. Nichols, "Surgery digitized: Telesurgery becoming a reality", *ZD Net,* 14 June 2021. Available at: https://www.zdnet.com/article/surgery-digitized-telesurgery-becoming-a-reality/.

[113] Kenneth, 2018, p. 233.

[114] Heaven, 2021. See *infra* for a myth buster definition of Artificial General Intelligence (AGI).

[115] The question will be addressed in depth in Chapter 5.2. and 8.2. For the time being, it can be stated that the lack of a universal technical definition of AI systems directly affects understanding whether AI could be considered as agents of crimes.

[116] F. Macagno, "Definitions in Law", *Bulletin Suisse de Linguistique Appliquee,* 2010, p. 201.

[117] See S. Legg & M. Hutter, "A Collection of Definitions of Intelligence", *arXiv:0706.3639v1*, 2007 for a survey of AI definitions conducted up to 2007. For a collection of more recent definitions, see *infra.*

[118] The definition adopted is contained in Article 3 of the "*Proposal for a regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*", COM(2021) 206 final, 21 April 2022: "artificial intelligence system' (AI

pertaining to computer science. Undeniably, "[a] complicating factor is that legal definitions differ from pure scientific definitions whereas they should meet a number of requirements (such as inclusiveness, preciseness, comprehensiveness, practicability, permanence), some of which are legally binding, and some are considered good regulatory practice".[119] This is especially true when it comes to criminal norms, which are subject to a strict principle of legality.

Having acknowledged that defining AI is indeed an issue, we need to ask ourselves why this research needs to adopt a definition. The reasons are manifold. To begin with, if we look at the concept of criminal liability one of its founding bricks is the concept of *action*: in order to understand how an AI acts, and why it does so, we need to first define what we mean by AI and who the artificial agent is. As it will be shown in Chapter 2, the lack of a common definition impacts directly on identifying criminally capable subjects, i.e., the addressees of criminal norms. Once this is established, it is possible to discuss relevant questions such as: *is AI behavior relevant for criminal law? What is the difference – if any – between the act of an AI and of a human being from a criminal law standpoint?* As a matter of fact, AI "[f]irst […] invites us to consider whether AI agents are acting in the sense of criminal law. And secondly, it urges us to think about different modes of acting when it comes to human agents".[120]

To continue, things seem to get even more complicated with regards to *mens rea* requirements: if AI entities are able to distinguish from right and wrong, what is permitted from what is forbidden, one could ask herself whether such capabilities can be seen as 'clues' of the existence of the internal elements needed for criminal liability. These and more aspects will be analyzed in depth in Chapters 4 and 6.

---

system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with". Annex I reads:

> (a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
> (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
> (c) Statistical approaches, Bayesian estimation, search and optimization methods.

[119] Council of Europe, CAHAI Secretariat, "Towards a regulation of AI systems", DGI (2020) 16, p. 23. Available at: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.

[120] Lima, 2018, p. 681.

The purpose of this Chapter is twofold: first, to adopt a working definition of AI and second to function as a primer for the reader, providing her with an adequate "toolbox" of notions that will be used throughout this analysis. The Chapter will be structured as follows. First, the author will give a synopsis of the development of AI accounting for approximately the past fifty years. Second, the focus will shift onto discussing the main definitions of AI that have been developed by experts in the field. Third, basic notions of current AI technologies will be provided. Fourth, the author will enucleate the adopted working definition for the research. Fifth, the author will provide real life examples of AI systems, which can be used as case studies. Finally, the Chapter will be concluded with an essential glossary of terms.

2.2 A SHORT HISTORY OF AI

When analyzing the history of AI, it is common to find analogies with the seasons to describe the cycles of hype and disappointment that characterized its development.[121] Generally, in the scholarly discourse identifies four distinct "springs" and "winters": the first spring (1956–1974), the first winter (1974– 1981), the second spring (1981–1987), and the second winter (1987–1993) (Fig. 2).[122] The use of seasonal metaphors is sometimes troublesome, as it reinforces a narrative which alternates cyclical portrays of AI as the "ultimate panacea, which would solve everything and overcome everything; or as the final catastrophe, a superintelligence that would destroy millions of jobs, replacing lawyers and doctors, journalists and researchers, truckers and taxi drivers, and ending by dominating human beings as if they were pets at best".[123]

---

[121] M. Haenlein & A. Kaplan, "A Brief History of Artificial Intelligence: On the Past, Present and Future of Artificial Intelligence", *California Management Review*, Vol. 61, No. 4, pp. 5–14, 2019; S. J. Russell & P. Norvig, *Artificial Intelligence. A Modern Approach*, 2nd edition, Pearson, 2003, pp. 16-27.

[122] Y. Shin, "The Spring of Artificial Intelligence in Its Global Winter", *IEEE Annals of the History of Computing,* Vol. 41, Issue 4, 2019.

[123] L. Floridi, "AI and Its New Winter: from Myths to Realities*",* in *Philosophy & Technology,* Vol. 33*,* 2020, p. 1.

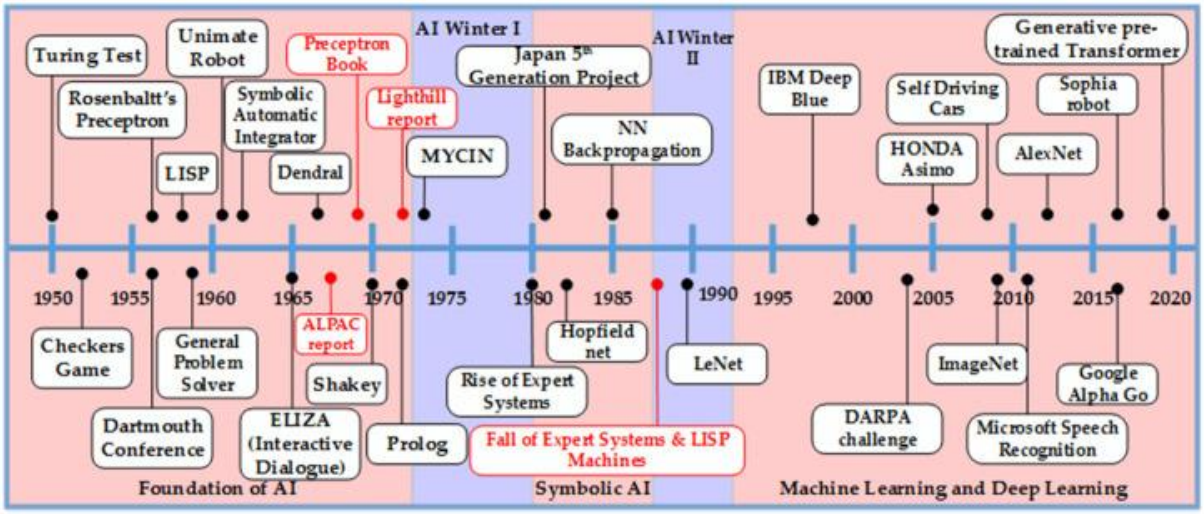

Figure 2 Artificial intelligence over the years. Source: F. H. Khan et al., "Advancements in Microprocessor Architecture for Ubiquitous AI—An Overview on History, Evolution, and Upcoming Challenges in AI Implementation", Micromachines, *Vol. 12, No.6, 2021.*

The official nascence of AI as a new scientific discipline is commonly seen in the year 1956.[124] For many, however, it can be located earlier in the work of Alan Turing in the 1940s. Notably, it was "[t]he powerful way in which The Bombe was able to break the Enigma code, a task previously impossible to even the best human mathematicians" [125] which made Turing wonder about the intelligence of said machine. In 1950, Turing published "Computing Machinery and Intelligence", in which he introduced the "imitation game", later known as the "Turing test".[126] Its goal was to establish whether a machine was intelligent (or not) based on the assumption that if a human interacting with an artificial intelligence was incapable of distinguishing it from a human, then the machine must be intelligent.

In the first wave of AI development, which was roughly the twenty years after 1956, research focused on 'symbolic' AI: humans pre-programmed systems to solve tasks by feeding them logical 'if-then' rules in the form of symbols (such as graphs or formulas). Symbolic AI requires "human experts to encode their knowledge in a way the computer can understand" and this "places significant constraints on their degree of autonomy".[127]

---

[124] McCarthy et al., 1955.

[125] Haenlein & Kaplan, 2019, p. 6.

[126] A. Turing, "Computing Machinery and Intelligence", *Mind*, LIX/236, 1950, pp. 433-460.

[127] European Parliament, "How artificial intelligence works", Briefing, Member's Research Service, 2019. Available at:

It was an extremely prosperous time for AI, one characterized by heavy investments from national governments.[128] One of the most famous events of that time was the introduction of ELIZA, a natural language processing computer program which mimicked a psychotherapist and that could "carry a conversation in English on any topic";[129] another was the introduction of the General Problem Solver, a computer program which could supposedly simulate human problem solving.[130] The optimism was palpable: AI researchers predicted that machines would soon become world chess champions and be able discover new mathematical theorems. Minsky in a famous 1970 Life Magazine interview professed

> In from three to eight years we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office politics, tell a joke, have a fight. At that point the machine will begin to educate itself with fantastic speed. In a few months it will be at genius level and few months after that its powers will be incalculable.[131]

Those predictions did not come true, mostly due to the fact that the machines had very little memory and the computing time was too long. What is more, symbolic systems suffer from the so-called "Moravec's paradox", which entails that it is easy to create computers which exhibit adult level performance on intelligence tests (such as the Turing test or playing chess), but it is difficult or impossible to give them the skillset of a one-year-old when it

---

https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/634420/EPRS_BRI(2019)634420_EN.pdf.

[128] Think for example of the $ 2.2 million grant from the Advanced Research Projects Agency (U.S.) to the MIT to fund the project MAC founded by Minsky and McCarthy.

[129] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine", *Communications of the ACM*, Vol. 9, Issue 1, 1966, pp. 36–45. A JavaScript version of Eliza is available at: http://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm.

[130] See H. A. Simon, J. Shaw & A. Newell, "Heuristic Problem Solving: The Next Advance in Operations Research", *Operations Research*, Vol. 6, No. 1, 1958; H. A. Simon, J. Shaw & A. Newell, "Report on a General Problem-Solving Program", *Rand Corporation*, 1959.

[131] B. Darrach, "Meet Shaky, the first electronic person", *Life Magazine*, 20 November 1970, p. 58B. Available at: https://books.google.it/books?id=2FMEAAAAMBAJ&lpg=PA57&dq=%22first%20electronic%20person%22&pg=PA58#v=onepage&q=years&f=false.

comes to perception and adaptability.[132] To put it simply: why is AI so smart and yet so dumb? Why does AI struggle with the simple?[133] Such questions, as it will be seen, are still relevant today.

A new stream of techno-optimism resurged at the end of the 1970, as symbolic artificial intelligence led to the development of "expert systems"(Fig.3). Such term referred to the belief that machines could reproduce the steps of a human expert when performing an assignment: the underlying idea was that intelligent human behavior could be deconstructed into a succession of logical rules, which could then be transcribed into algorithms that machines could follow (according to a "top down" approach).[134]

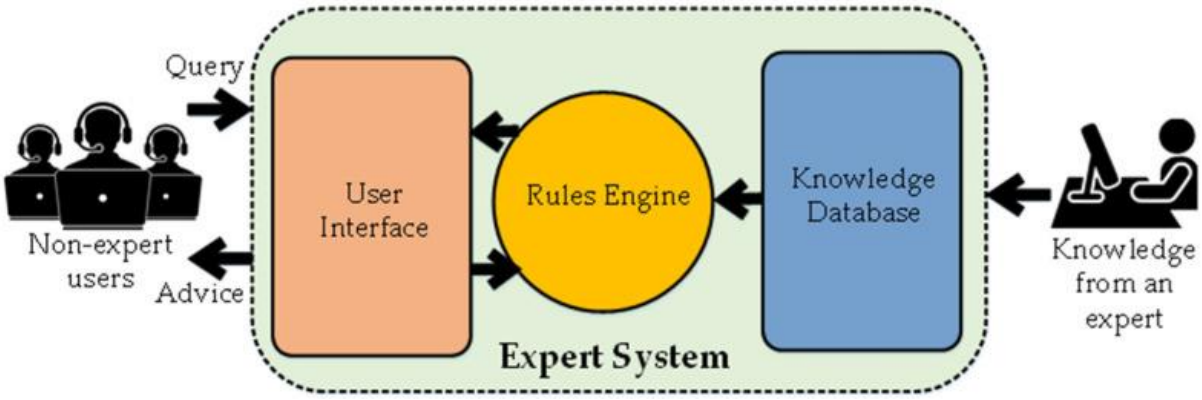

Figure 3 Expert Systems. Source: Khan, 2021.

Expert systems, though, had limits. One of these is that they can only know as much as humans could teach them. Likewise, they are not flexible: they cannot adapt to problems where both the variables and the rules change in time, hence they work best in highly formalized areas, such as chess.

---

[132] H. Moravec, *Mind Children. The future of Robot and Human Intelligence*, Harvard University Press, 1990, p. 15.

[133] R. C. Suwandi, "Why is AI So Smart and Yet So Dumb? What Moravec's Paradox told us about AI", Towards data science, 30 august 2020. Available at: https://towardsdatascience.com/why-ai-is-so-smart-and-yet-so-dumb-c156cc87fafa.

[134] European Parliament, "Understanding Artificial Intelligence", Briefing, Member's Research Service, 2018. Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/614654/EPRS_BRI(2018)614654_EN.pdf.

In 1997, IBM's supercomputer Deep Blue[135] defeated the world chess champion Garry Kasparov. Chess was considered "the pinnacle of human intelligence".[136] Yet, this result alone did not boost confidence (nor funding) in AI: Deep Blue had won by using "brute force computing power",[137] i.e., its ability to examine 200 million chess moves per second. It was thanks to its speed and capacity that it had beaten Kasparov, not thanks to its "intelligence". As it was argued, chess does not represent "the crowning glory of human intellectual endeavour; it is simply a mathematical problem with very clear rules and a finite set of alternatives".[138]

What brought AI to a new era was a paradigm change: the age of data-driven artificial intelligence,[139] where is no longer about coding rules for expert systems, rather, it is about allowing computers to sift enormous amounts of data to find correlations and classifications.[140] The second wave of AI development, then, is based on two elements: first, the enormous quantity of data available at little or no cost; second, the development of algorithms capable of learning by themselves (hence not through programmers encoding them with their previous knowledge) through machine learning techniques (ML). Such techniques had been available for some time, yet their advance surged ahead thanks to the increase in the availability of digital data.[141]

What about today? As it will be shown, the type of AI that is currently deployed poses fundamental questions to traditional notions of criminal liability. That will be the object of this inquiry, no matter the season.

---

135 IBM, "Deep Blue". Available at https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/. See also: M. Campbell, A. J. Hoane Jr. & F. Hsu, "Deep Blue", *Artificial Intelligence*, Vol. 134, 2002, pp. 57–83.

136 H. Sheikh, C. Prins, & E. Schrijvers, *Mission AI. Research for Policy,* Springer, 2023, p. 17.

137 H. Yu, "From Deep Blue to DeepMind: What AlphaGo Tells Us", *Predictive Analytics and Futurism*, Issue 13, 2016, p. 43 See also Council of Europe, "History of Artificial Intelligence". Available at: https://www.coe.int/en/web/artificial-intelligence/history-of-ai; G. Press, "The Brute Force Of IBM Deep Blue And Google DeepMind", *Forbes online*, 7 February 2018. Available at: https://www.forbes.com/sites/gilpress/2018/02/07/the-brute-force-of-deep-blue-and-deep-learning/?sh=597fb30249e3.

138 Sheikh, Prins & Schrijvers, 2023, p. 17.

139 European Parliament, "Understanding Artificial Intelligence", 2018.

140 Council of Europe, "History of Artificial Intelligence".

141 Council of Europe & Yeung, 2019, p. 17.

2.3 GAINING A BASIC UNDERSTANDING OF MODERN AI SYSTEMS

Since most AI systems mentioned in this research are based on ML, it is relevant at this point to explain its functioning and its applications.

An algorithm based on ML techniques teaches itself rules by learning from the training data and through statistical analysis, detecting patterns in large amounts of information and generating outputs. It adopts a bottom down approach, unlike knowledge-based systems. These patterns can then be applied in different tasks, such as driving a car.[142] ML algorithms can be trained through supervised or an unsupervised learning, depending on whether the algorithm learns from previously labelled data or not.

How does machine learning work? Relying on an analogy, one could say that it is similar to learning how to ride a bike: "[y]ou don't tell a child to move their left foot in a circle on the left pedal in the forward direction while moving their foot in a circle… You give them a push and tell them to keep the bike upright and pointed forward: the overall objective. They fall a few times, honing their skills each time they fail".[143]

A highly important subset of ML is deep learning (DL), which consists of layers of artificial neural networks (ANNs). Neural networks' engineering was inspired by the functioning of biological neurons: their basic function is to establish features from an input. ANNs are made of layers of functions ("nodes" or "neurons") that perform various operations on the data that they are fed (Fig. 4). The more layers a network has, the "deeper" the deep learning is.[144]

---

[142] Surden, 2019, p. 1311.

[143] T. Havens quoted in K. Casey, "How to explain machine learning in plain English", *The Enterprisers Project*, 19 November 2020. Available at: https://enterprisersproject.com/article/2019/7/machine-learning-explained-plain-english?page=0%2C0.

[144] C. Shenkman, D. Thakur & E. Llansó, Center for Democracy & Technology, *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis*, 2021, p. 43. Available at: https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf.

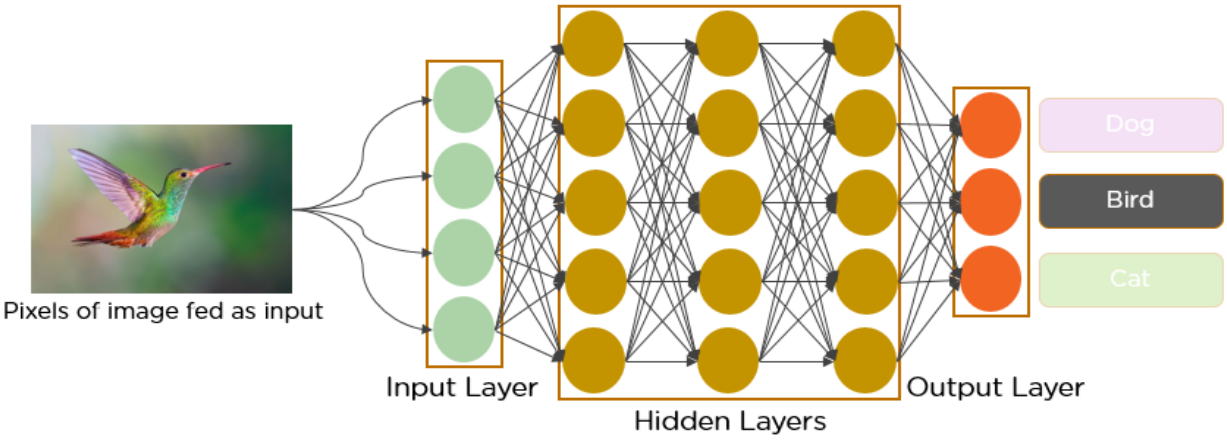

Figure 4 Functioning of an Artificial Neural Network. Source: M. Mandal, "Convolutional Neural Networks (CNN)", Medium, *30 April 2021. Available at: https://manavmandal.medium.com/convolutional-neural-networks-cnn-d88e7f9329eb.*

The difference between ML and DL is that in DL the algorithm is fed raw (unlabeled) data and then it identifies which features are relevant. In ML learning, instead, the algorithm is given an established set of relevant features to analyze.

If we rely once again on the analogy with a child, we could think of DL as a child learning language. Imagine a child that points at an object and says the word "car". The child's parents will immediately provide a feedback such as "right" or "wrong" or provide a different label for the object ("No, that's a car!"). After a certain amount of feedback, the child will form a mental model of how to correctly label objects that she perceives in the real world. This is accomplished thanks to neurons that transmit signals to other neurons, "some sort of unexplainably complex hierarchy that is formed based on feedback". [145]

To understand the relationship between AI, ML, ANNs, and DL one should think of them like Russian nesting dolls, where each concept is part of the previous (Fig.5). Viewed this way, it is easier to grasp that machine learning is a subfield of artificial intelligence, deep learning is a subfield of machine learning, and neural networks "make up the backbone of deep learning algorithms […]".[146] Based on the number of node layers or depth of neural

[145] Casey, 2020.

[146] E. Kavlakoglu, "AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?", *IBM*, 27 May 2020. Available at: https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks.

network it is possible to distinguish a single neural network from a deep learning algorithm, "which must have more than three".[147]

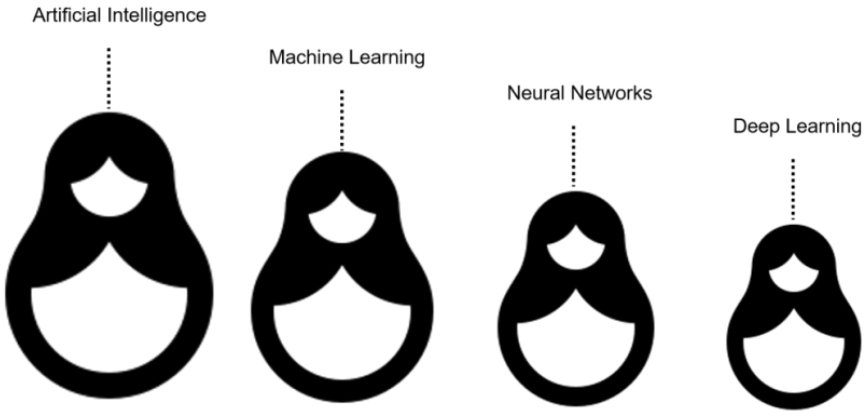

Figure 5. Relationship between AI, ML, ANNs and DL. Source: Kavlakoglu, 2023.

One of the DL methods that is mostly used for solving tasks involving image classification (a core task for AVs, but also for robotics applications)[148] is Convolutional Neural Networks (CNN). CNNs are multi-layered neural networks whose creation was inspired by animal optical systems, which are based on visual cortex cells that detect light in the small receptive field.[149] CNNs extract features from images, identify objects that are contained in them and then classifies the images (i.e., general object recognition task).

Let us walk through these concepts by looking at an oversimplified example. Imagine that the task that has to be solved by the system is to classify whether an image contains a pedestrian, another vehicle or a traffic light. Such an operation is useful, for example, when developing AVs in order to avoid accidents (Fig. 6).[150] CNNs can be thought of as a very large car factory where the process of establishing the output (classifying a certain object in a picture as a human being/not a human being) is broken into a multitude of sub-tasks performed by millions of highly skilled (pre-trained) workers divided into teams (the "nodes").

---

147 Ibid.

148 I. Sikdokur, I. Baytas & A. Yurdakul, "Image Classification on Accelerated Neural Networks", *arXiv:2203.11081v1* [cs.CV], 2022.

149 F. Sultana, A. Sufian & P. Dutta, "Advancements in Image Classification using Convolutional Neural Network", *arXiv:1905.03288v1* [cs.CV], 2019.

150 This technique is called "Computer Vision" and it comprises of DL models and CNNs.

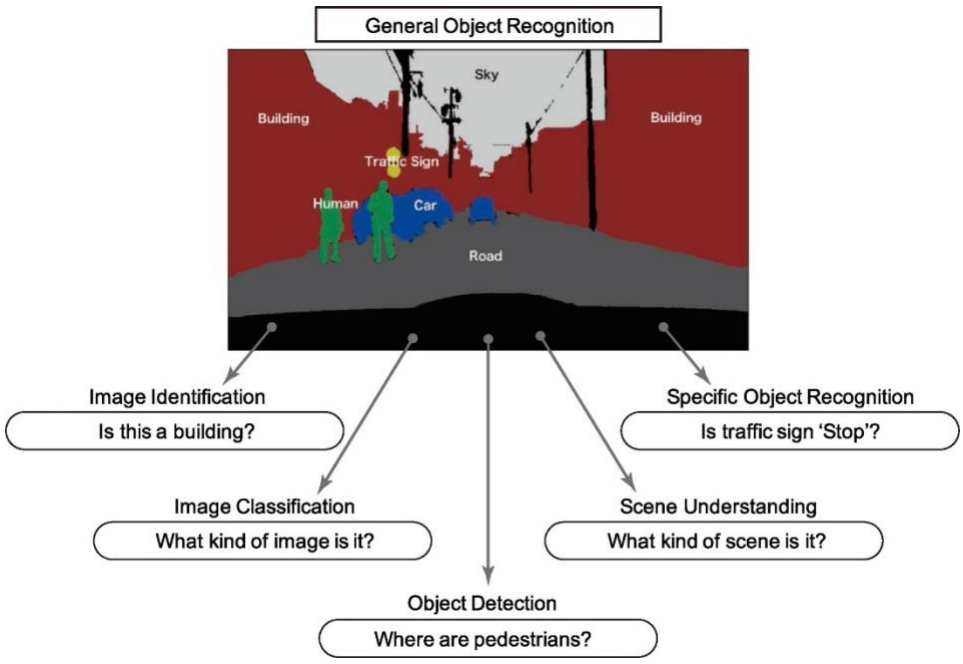

Figure 6. Segmentation of general object recognition. Source: H. Fujiyosh, T. Hirakawa & T. Yamashita, "Deep learning-based image recognition for autonomous driving", IATSS Research, Vol. 43, Issue 4, 2022.

Imagine now that the first team is given a very large dataset of pictures. The workers will apply a filter to the input image to identify certain features, for example, the perimeter of the object and then pass this output to the next team, which based on this output will identify legs or arms (and so on) until the final output is the classification of the object in the picture. The work of one team builds on the output of the previous team, though, and such "intermediate" output might look nothing like the finished product ("much as an ignition coil may not be immediately recognizable as a car part",[151] even to expert car drivers). Moreover, during the process "tasks and workflow may also be shifted in real-time to make the process more efficient".[152]

One fundamental issue with DL, which will often reappear in the following Chapters, is that it is "inherently a black box": [w]hile it is straightforward to assess the quality of the output generated by such systems (e.g., the share of correctly classified pictures), the process used for doing so remains largely opaque. Such opacity can be intentional (e.g., if a

[151] This example is an adaption of the one contained in Shenkman, Thakur & Llansó, 2021, p. 43.
[152] Shenkman, Thakur & Llansó, 2021, p.43.

corporation wants to keep an algorithm secret), due to technical illiteracy or related to the scale of application (e.g., in cases where a multitude of programmers and methods are involved).[153] In other words, someone walking through the car factory "would likely find it impossible to grasp the immensity of the process or the relationships between various teams and processes".[154]

The second fundamental issue, which acquires relevance from a criminal law perspective, is the 'autonomy' property of systems based on ML. Currently employed AI systems always work to fulfill an overarching goal set by a human, yet they can decide independently between alternatives ways to reach said goal, and they can set their own intermediate sub-purposes.[155]

On a last note, it is relevant to briefly touch upon the theme of hybrid AI systems. This term is used to refer to two types of hybridization: on the one hand, AI systems can comprise of different techniques, such as machine learning combined with knowledge-based systems.[156] On the other hand, AI systems can involve forms of human decision-making in their functioning. This concept will prove particularly significant for this investigation of the realm of criminal liability. Human-AI interaction can happen in different ways: training, tuning testing and operating the model.[157] In this regard, three main approaches are important: human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC). HITL is an umbrella term which refers to the ability for the human agent to intervene in every decision of the system.[158] More specifically, HOTL, instead, refers to "the capability for human intervention during the design cycle of the system and monitoring the system's operation". [159] Finally, HIC refers to the ability of the human to oversee the

---

153 Haenlein & Kaplan, 2019, p.1.

154 Shenkman, Thakur & Llansó, 2021, p. 43.

155 Council of Europe & Yeung, 2019, p. 19.

156 K. Martineau, "Teaching machines to reason about what they see", *MIT News*, 2 April 2019. Available at: https://news.mit.edu/2019/teaching-machines-to-reason-about-what-they-see-0402; W. Knight, "Two rival AI approaches combine to let machines learn about the world like a child", *MIT Technology Review*, 8 April 2019. Available at: https://www.technologyreview.com/2019/04/08/103223/two-rival-ai-approaches-combine-to-let-machines-learn-about-the-world-like-a-child/.

157 See also F. M. Zanzotto, "*Human-in-the-loop Artificial Intelligence*", *Journal of Artificial Intelligence Research*, Vol. 64, 2019, pp. 243-252.

158 AI-HLEG, *Ethics guidelines for trustworthy AI*, 2019, p. 16. Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

159 Ibid.

overall activity of the system (which includes also its broader economic, societal, legal and ethical impact) together with the ability to decide "when and how to use the system in any particular situation".[160]

Operating a human-in-the-loop system is diffused when it comes to Autonomous Vehicles ("AVs" from this point forward): currently, most private companies that operate AVs are required by law to have a human driver present in the car.[161] The role of the human driver is to reassert full control over the car when an error happens or to avoid an accident, as a sort of duty of care.[162] What happens, then, if the driver was not able to take over in a timely manner and a deadly accident occurs?[163] One should think of the case of an elderly driver, which is often labelled a dangerous driver when it comes to "normal" cars and is therefore supposedly a safer driver when traveling in an AV. The same person, though, is likely to be less skilled concerning reestablishing digital control over the AV. Should she be blamed for the death of a pedestrian because she did not assume control of the car in time? As it will be explained in Ch. 5, the UK Law Commissions addressed this issue and suggested the introduction of a new law providing for an immunity clause for drivers (called users-in-command") against offenses committed by self-driving vehicles.

---

[160] Ibid.

[161] See for example the amendment to article 8 of the United Nations 1968 Vienna Convention on Road Traffic which provides that automated driving technologies transferring driving tasks to the vehicle will be explicitly allowed in traffic provided that these technologies are in conformity with the United Nations vehicle regulations or can be overridden or switched off by the driver. United Nations Economic and Social Council, *Report of the sixty-eighth session of the Working Party on Road Traffic Safety*, ECE/TRANS/WP.1/145, 2014. Available at: https://unece.org/press/unece-paves-way-automated-driving-updating-un-international-convention). See also Article 3 of the United Kingdom Automated and Electric Vehicles Act of 2018 which provides that "The insurer or owner of an automated vehicle is not liable under section 2 to the person in charge of the vehicle where the accident that it caused was wholly due to the person's negligence in allowing the vehicle to begin driving itself when it was not appropriate to do so". Automated and Electric Vehicles Act, 2018, Chapter 18.

[162] B. Wagner, "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems", *Policy & Internet*, Vol. 11, Issue 1, 2019, p. 109.

[163] For an analysis of the development of shared operational responsibility between vehicle and driver, see J. Pattinson, C. Haibo & B. Subhajit, "Legal issues in automated vehicles: critically considering the potential role of consent and interactive digital interfaces", *Humanities and Social Sciences Communications*, Vol. 7, Art. No. 153, 2020.

More difficulties arise when an AV commits an error which is different from "common" human driving errors and, as such, hard to predict for the human driver.[164] We can think of the case where a car in broad daylight and perfect visibility is faced with the option of hitting a person or a very roughly made scarecrow: in this case, the failure of the car's sensors to identify a figure as human is radically different from not stopping at a red traffic light. Can the driver be considered at fault for not anticipating such a "weird" action of the system?

Why are these notions relevant to this research? Recently, it has been advocated especially in policymaking that adopting a HITL-approach is the ideal solution to ensure accountability: "if an autonomous system causes harm to human beings, having a human in the loop provides trust that somebody would bare the consequence of such mistakes".[165] Researchers have thus introduced the theory of the "big red button", a sort of emergency kill switch for AI systems, which should stop it before it becomes destructive.[166] Such mechanisms, as has been highlighted through the AV example, need to be handled with care, especially if they are to be translated to criminal liability.[167] These issues are addressed in greater detail in the discussion that will be conducted in Chapter 6.

---

164 "Machine errors in automated driving are typically associated with the correct identification of the objects perceived by sensors, while human driving errors are typically associated with being able to provide sustained attention to a specific task over long periods of time". Wagner, 2019, p. 109.

165 European Parliamentary Research Service, Panel for the Future of Science and Technology, "The ethics of artificial intelligence: Issues and initiatives", EPRS_STUD(2020)634452, 2020, p. 35. See also M. Pizzi, M. Romanoff & T. Engelhardt, "AI for humanitarian action: Human rights and ethics", *International Review of the Red Cross*, Vol. 102, No. 913, 2020, pp. 145–180; AI-HLEG, 2019.

166 T. Arnold & M. Scheutz, "The "big red button" is too late: an alternative model for the ethical evaluation of AI systems", *Ethics and Information Technology*, Vol. 20, 2018, p. 60.

167 "Accountability mechanisms built on the assumption of a supreme human overseer are inherently flawed, if adopted without criticism. Such approaches can embed and reinforce the implicit human/machine dichotomy and mystify human agency. However, it should be noted that the emphasis on human agency might serve a purpose outside monitoring automation, namely in justifying legal decisions. The importance attributed to humans in automation is not arbitrary but instead reflects the legal system's foundational concepts and ideologies that are built on anthropocentricity. In other words, juxtaposing algorithmic and human decision-making reveals law's self-reflection on what constitutes legal decision-making. Simply put, law's acknowledgement of legal agents capable of decisions is limited to humans or fictions of human agents such as organizations that are conceptualized as legal (although not natural) persons. Following this, justification of decision-making has traditionally been connected to human agency even when, in practice, decisions are arrived at through intra-organizational processes. In this sense, human control over automation can be seen simply as another formulation of human justification." R. Koulou, "Human Control over Automation: EU Policy and AI Ethics", *EJLS* 12(1), 2020, pp. 9-46.

2.4 DEFINITIONS OF AI: AN OVERVIEW

As it was shown, thinking about AI has changed over time.[168] It was first defined by John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon in 'A proposal for the Dartmouth Summer Research Project on Artificial Intelligence', a "2 month, 20 man study" which took place in 1956.[169] They defined it as "making a machine behave in ways that would be called intelligent if a human were so behaving".[170] Fifty years later, McCarthy published the article "What is artificial intelligence?" in which he defined AI as "the science and engineering of making intelligent machines, especially intelligent computer programs".[171] It can be argued, then, that AI is "first and foremost a *science*, classified in a variety of subfields, whose general aim is that of creating intelligent machines".[172] This can be referred to as an 'operational definition': defining AI by what AI researchers do.[173] Such subfields or technologies include machine learning, deep learning, computer vision and natural language processing.

An added difficulty to defining AI in these terms is that the public discourse suffers from the so-called "AI effect" or "odd paradox". These terms describe the phenomenon of AI reaching mainstream usage and then not being considered AI anymore, as people become

---

168 K. J. Hayward & M. M. Maas, "Artificial intelligence and crime: A primer for criminologists", Crime Media Culture, 2020, p. 4. For a systematic account of policy making definitions of AI adopted by national governments and international organizations, see A. Bertolini, "Artificial Intelligence and Civil Liability", Study commissioned by the European Parliament's Committee on Legal Affairs, European Union, IPOL_STU(2020)621926, 2020, pp. 23-29. For a thorough overview of existing AI definitions, see the research conducted on 55 documents by S. Samoili et. al, "AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence", EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, JRC118163.

169 McCarthy et al., 1955.

170 Ivi, p. 12.

171 J. McCarthy, "What is artificial intelligence", 2007, p. 2. Available at: http://jmc.stanford.edu/articles/whatisai.html.

172 Bertolini, 2020, p. 18.

173 Stone P. et al., "Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel", Stanford University, September 2016. Available at: http://ai100.stanford.edu/2016-report.

accustomed to it.[174] Shortly put, "[a]s soon as it works, no one calls it AI anymore".[175] Moreover, drawing on an often-cited analogy in the field of AI-liability, AI systems cannot define themselves and in this they differ profoundly from corporations, which define themselves as such once they are established and acquire legal personality.

Additionally, a further problematic aspect of defining AI from a regulatory perspective is to outline the *object* of this field of science,[176] namely defining what intelligence is, be it human or artificial. McCarthy notes that intelligence entails the ability of achieving goals in the world.[177] Various degrees and kinds of intelligence can occur in human beings, animals and in some machines: the issue is that, since it is not possible to characterize (yet) what kinds of computational procedures we want to call intelligent, we cannot answer definitely on whether a machine is intelligent or not (yet).[178]

One can distinguish two categories of definitions in this sense: rationalist definitions and human centric definitions. Rationalist definitions focus on the ideal concept of intelligence, referred to as "rationality".[179] Norvig-Russel's *Artificial Intelligence: A Modern Approach*, one of the most distributed textbooks on AI, adopts said approach and defines AI as the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment.[180]

Human centric definitions, on the other hand, define AI in terms of fidelity to human performance (i.e. systems that think and act like humans). Examples of human centric definitions include describing AI as the study of how to make computers do things at which at the moment people are better[181] or as the study of computers doing tasks that would be

---

174 Stone et al., 2016, p. 12.

175 J. McCarthy quoted in M. Y. Vardi, "Artificial Intelligence: Past and Future", *Communications of the ACM*", Vol. 55, No 1, 2012, p. 5.

176 Bertolini, 2020, p. 20.

177 McCarthy, 2007, p.2.

178 Ibid.

179 I. Ben-Israel et al., "Towards Regulation of AI Systems. Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law", Council of Europe Study DGI (2020) 16, 2020, p. 22.

180 Norvig & Russel, 2003, pp. 1-2.

181 E. Rich & K. Knight, *Artificial Intelligenc*e, Tata McGraw, 2004, p. 3.

considered to require intelligence if a human did them (tasks that "normally require human intelligence").[182]

As a final note, the AI systems that are employed today are "narrow", meaning that they can only solve pre-established tasks. In other words, today's AI is not *actually* intelligent.[183] As it has been suggested, "[i]t may even be proposed, as a rule of thumb, that any activity computers are able to perform and people once performed should be counted as an instance of intelligence. *But matching any human ability is only a sufficient condition, not a necessary one*. There are already many systems that exceed human intelligence, at least in speed, such as scheduling the daily arrivals and departures of thousands of flights in an airport".[184]

Is the comparison to human intelligence then truly needed for this research? As argued by Armer in his "Argument of Invidious Comparison",

> Considering the behavior of men and machines in the context of intelligence being a multidimensional continuum is like saying that the Wright brothers' airplane could not fly because it could not fly nonstop from Los Angeles to New York nor could it land in a tree like a bird. Why must the test of intelligence be that the machine achieve identically the same point in the continuum as man? Is the test of flying the achievement of the same point in the continuum of flying as that reached by a bird?[185]

Stretching this even further, AI can cause damage, without intentionally wanting to cause it. What matters, then, is the potential for harm, irrespective of whether one is discussing an algebraic formula or a deep neural network.

In 2015, AlphaGo, developed by Google, beat the world Go Champion Lee Sedol. As some have recognized "AlphaGo mimics an extremely diligent, but not necessarily genius, student who is willing to learn from millions of human's play and self-play, tediously."[186] In

---

[182] M. L. Minsky quoted in T. Stonier, *Beyond Information. The Natural History of Intelligence*, Springer, 1992, pp. 107-133.
[183] Surden, 2019, p. 1308.
[184] Stone P. et al., 2016, p. 13.
[185] "Why must the test of intelligence be that the machine achieve identically the same point in the continuum as man? Is the test of flying the achievement of the same point in the continuum of flying as that reached by a bird?". P. Armer, "Attitudes Toward Intelligent machines", in E. A. Feigenbaum, J. Feldman & P. Armer (Eds.), *Computers and Thought,* Aaai Pr, 1995, p. 393.
[186] Yu, 2016, p. 44.

fact, in today's world, "[p]laying chess against a machine—and losing with near certainty—has become a thing not even worth mentioning".[187] Being able to play chess is not taken as the yardstick of intelligence anymore, nor is being able to play Go. Without any doubt, former chess champion Garry Kasparov had an entirely different view on this matter in 1997, after his defeat against Deep Blue.[188] So did society at the time: think of the cover of Newsweek of 5 May 1997 which was titled "The Brain's Last Stand".

2.5 ADOPTED WORKING DEFINITION

This study will adopt as a working definition the HLEG definition of AI:

> Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal.
>
> AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.
>
> As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).

[187] Haenlein & Kaplan, 2019, p.13.
[188] Haenlein & Kaplan, 2019, p.13.

The reasons for this choice are numerous. First of all, it is a comprehensive definition which comprises of the main features of AI according to the most recurring definitions in the academic discourse: perception of the environment, information processing (collecting and interpreting inputs in data), decision making (including reasoning and learning), the performance of actions and tasks in adaptation (in reaction to changes in the environment) with a certain level of autonomy; achievement of specific goals.[189] Secondly, it avoids ambiguities regarding the definition of "intelligence" as it does not define AI as an 'intelligent' machine. Such a definition will guide this research and represents a compromise with different definitions.[190] The term AI will be used when referring to artificial intelligence as a type of technology. The term AI systems will instead be adopted when referring to AI as an application of the technology.[191]

From the abovementioned definition one can identify at least two problematic characteristics that are closely related to the functioning of AI-systems: unpredictability and autonomy. The focus of this research will be only on those AI systems that are programmed in such a way that they become adaptive or self-deciding agents, i.e. whose actions are not *ex ante* open and predictable.[192] Said issues will be addressed in the relevant Chapters.

2.6 CONCLUSIONS

Some have said that "[d]efining AI is an exercise rather like nailing jello to a tree: with forethought, planning, and enough nails it ought to be doable, but it isn't".[193] Let us not be

189 Samoili et al., 2020, p. 8.

190 P. Wang, "What Do You Mean by 'AI'?", *Proceedings of the 2008 conference on Artificial General Intelligence*, 2008, pp. 362–373.

191 Scherer M.U., "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies", *Harvard Journal of Law & Technology*, Vol. 29, No. 2, 2016, p. 362.

192 "Deshalb sei vorausgesetzt, dass hier nur solche „Roboter", „Maschinen" oder „Software-Agenten" von Interesse sind, die so programmiert sind, dass sie empfangene Informationen in einer Weise verknüpfen und für künftige Reaktionen auswerten können, die nicht vollständig durch die Programmierung festgelegt ist und deren Aktionen daher ex ante „offen" und mithin nicht für jeden Einzelfall vorhersehbar sind. Nur hinsichtlich solcher „lernfähiger" bzw. „selbst entscheidender" Agenten kann überhaupt die Frage gestellt werden, ob ihnen das Strafrecht eine eigene, personale Rolle zuweisen kann. Alle anderen, „fest" programmierten Geräte können – so will ich es unterstellen – nur als „Werkzeuge" in der Hand menschlicher Agenten in Betracht kommen". Seher, 2016, p. 46.

193 D. Partridge, *A New Guide to Artificial Intelligence*, Intellect L & D E F A E, 1991, p. 1.

discouraged: this Chapter does not have the (unattainable) goal of giving a final answer to the issue of defining AI. This research is meant to analyze AI through the lenses of classical notions of criminal law, therefore the purpose of this Chapter was to overcome the conceptual challenge by adopting a working definition. When doing so, it offered a basic characterization of the most relevant uses of AI. This provides the reader with a toolkit of notions adequate for addressing the implications of AI on the concept of criminal liability. Having gained the correct amount of technical knowledge we needed; it is possible to move to the next "trouble with AI": how to assign criminal responsibility for AI-related harm.[194]

[194] Scherer, 2016, p. 358.

3 EXPANSIONISTS, MODERATES, SKEPTICS: THE SCHOLARLY DEBATE ON AI AND CRIMINAL LAW

The civilian puts his faith in syllogisms, the common lawyer in precedents;
the first silently asking himself as each new problem arises, "What should we do this time?"
and the second asking aloud in the same situation, "What did we do last time?"
The civilian thinks in terms of rights and duties,
the common lawyer in terms of remedies.
The civilian is chiefly concerned with the policy and rationale of a rule of law,
the common lawyer with its pedigree.
The instinct of the civilian is to systematize.
The working rule of the common lawyer is solvitur ambulando.

Thomas Mackay Cooper, "'The Common and the Civil Law—A Scot's View",
63 Harv. L. R. 468 (1950).

3.1 Introduction – **3.2** Expansionists (or the Front of Robotic Liberation) – **3.2.1.** Gabriel Hallevy – **3.2.2.** Ying Hu – **3.2.3.** Christina Mulligan – **3.2.4.** Lasse Quarck – **3.3** Moderates – **3.3.1.** Ryan Abbott and Alex Sarch – **3.3.2.** Lagioia and Sartor – **3.3.3.** Freitas, Andrade, Novais and Osmani – **3.3.4.** Simmler and Markwalder – **3.3.5.** Mihail Diamantis – **3.4** Skeptics – **3.4.1.** The Italian Approach – **3.4.2.** Ugo Pagallo – **3.4.3.** Dafni Lima – **3.4.4.** Peter Asaro – **3.4.5.** The German Approach – **i.** Sabine Gless, Thomas Weigend and Emily Silverman – **ii.** Susanne Beck – **iii.** Gerhard Seher – **5** Conclusions

3.1 INTRODUCTION. THE SCHOLARLY DEBATE ON CRIMINAL LIABILITY OF AI-SYSTEMS

Up until ten years ago, academic pursuits to link criminal liability to the use of AI systems were wary. Most of the discussion and the policing efforts were limited to the areas of civil law and torts. In a 2020 research on the scope of legal literature on AI, [195] which was

[195] C. Rosca et al., "Return of the AI: An Analysis of Legal Research on Artificial Intelligence Using Topic Modelling", *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop*, 2020.

conducted with the aid of a machine learning technique called topic modelling,[196] researchers found that scholarly output boomed in the so-called "deep learning era".[197] As the authors argue, "with over 2500 publications already by the year 2015 referring to 'artificial intelligence' … it may no longer be realistic to assume that researchers can keep up with legal research on AI, or the number of publications in general".[198]

The scenario that researchers face nowadays with regards to the specific field of criminal legal scholarship is undeniably different: today AI and criminal law are no dichotomy. Indeed, as mentioned in the introduction, one of the purposes of the literature review is to situate this thesis in this "ocean" of scholarly literature.

Being late is no news for criminal law, and not necessarily a bad thing. Indeed, "[t]he criminal law is and ought to be different—importantly dissimilar from other kinds of law".[199] From a regulatory perspective, norms of criminal law represent "the ultima ratio in the legislator's toolkit".[200] One thing is certain: regardless of whether legislators will address criminal law issues pertaining to AI-related misbehaviors in the short run or not, starting from 2019 the academic discourse on AI in the realm of criminal law has thrived. The ultimate proof of this excitement can be found in the fact that the International Association of Penal Law decided to dedicate its XXIst International Congress in 2024 to the topic of "Artificial Intelligence in Criminal Justice".[201]

With this in mind, it is possible to appreciate why it is necessary that this research provides an overview of the current academic debate on the matter. Indeed, "[b]uilding your research on and relating it to existing knowledge is the building block of all academic research activities, regardless of discipline."[202]

---

[196] Topic Modelling is a method for classifying collections of documents. The authors adopted Latent Dirichlet Allocation (LDA) Topic Modelling. They used the tool to identify recurring topics in 3931 journal articles on AI legal research.

[197] The expression refers to the period which includes the early 2000s until today. C. Goanta et al., "Back to the Future: Waves of Legal Scholarship on Artificial Intelligence", in S. Ranchordás & Y. Roznai (Eds.), *Time, Law and Change*, Hart Publishing, 2020, p. 331.

[198] Rosca et al., 2020, p. 1.

[199] D. Husak, "The Criminal Law as Last Resort", *Oxford Journal of legal Studies*, Vol. 23, No. 2, 2004, p. 211.

[200] P. Minkkinen, "'If Taken in Earnest': Criminal Law Doctrine and the Last Resort", *The Howard Journal*, Vol. 45, No. 5, 2006, p. 526.

[201] See the information available at: http://www.penal.org/en/information.

[202] H. Snyder, "Literature review as a research methodology: An overview and guidelines", *Journal of Business Research*, Vol. 104(C), 2019.

Let us turn now to the purposes of this literature review. They are twofold. First, it establishes the conceptual framework for the research questions of the thesis and the main themes which will be addressed. By introducing the most relevant analyses of criminal legal scholars in relation to the technical concepts provided in Chapter 2, this section serves as foundation for the discussion contained in Chapters 4 and 5. In other words, this research is structured following a layered approach: at the beginning it provided the required background knowledge to be able, as lawyers, to understand foreign concepts, such as those pertaining to computer science; now it will deliver the reflections of the legal scholars that analyzed said concepts through the lenses of criminal law. Therefore, this Chapter will combine familiar notions, such as the ones pertaining to criminal legal theories, with unfamiliar notions, such as the ones pertaining to the realm of AI systems.

Second, it will serve to identify recurrent issues in the academic discourse on AI and criminal liability and, as a consequence, those that are often neglected. This will allow the research to, on one hand, compare different positions on the same issues and, on the other, acquire originality. The systematizing effort conducted in this Chapter is needed because of the panoply of approaches that are presented by the authors working on this topic: as if there were many musical instruments playing simultaneously, without a director to coordinate them.

This literature review suffers of few limitations. To begin with, due to the momentum that AI is living in criminal law at the time this research is conducted, it was not possible for this literature review to account extensively for each and any academic paper published in the field. Moreover, the analysis had to be limited to research published in three languages, namely English, Italian and German, as these are the languages spoken by the author.

The review was not limited with regards to the territorial provenance or the legal background of the authors, as the academic discourse in this field is characterized by intense crossbreeding: most of the papers discuss the impact of AI in criminal law by referring to general concepts of criminal law, rather than to how a specific concept is regulated in a national or international legal system. Authors frequently refer to works of other colleagues who have a different legal background from theirs and do so by mixing legal terms in English, German and Italian. This also rendered the operation of conducting the literature review difficult at times, due to the challenges represented by the transposition of concepts between different legal systems. For example, the Italian term "*colpevolezza*" is often used by Italian

scholars to refer to uses of the narrower concept of *mens rea*, without giving account of the difference between the two. Indeed, the concept of '*mens rea'* has gradually lost its broader meaning of 'blameworthiness' in American legal doctrine and is currently used primarily to refer to the specific mental state required by the criminal offense (e.g., intent).

On a final note, even though the legal background of the authors was not accounted for when selecting the relevant articles and book to read, it was addressed, if appropriate, when scrutinizing their content.

It is possible to identify three different streams of thought in the current debate on criminal liability of AI. As a consequence, the authors have been classified into three categories, "AI punishment expansionists", moderates and skeptics.[203] An explanation of these categories will follow in the appropriate subsection.

3.2 EXPANSIONISTS (OR THE FRONT OF ROBOTIC LIBERATION)

Expansionists or representatives of the Front of Robotic Liberation can be defined as those who argue for a radical change of the elements of criminal liability and/or for the direct punishment of AI systems. This label combines categorizations made by Abbott and Sarch[204] and by Pagallo.[205]

3.2.1 Gabriel Hallevy – the AI "Believer"206

Undoubtedly, one of the most cited legal scholars belonging to the expansionist front is Gabriel Hallevy.[207] He is the predecessor of almost all discussions on AI criminal liability (his

---

[203] Abbott & Sarch, 2019, p. 327.

[204] Ibid.

[205] Pagallo, 2013, p.54.

[206] U. Pagallo & W. Barfield, *Advanced Introduction to Law and Artificial Intelligence*, Edward Elgar Publishing, 2020, p. 140.

[207] Over the years, Gabriel Hallevy published several work comprising of both books and academic papers: "The Criminal Liability of Artificial Intelligence Entities - From Science Fiction to Legal Social Control, *Akron Intellectual Property Journal*, Vol. 4, Iss. 2, Art. 1, 2010; "I, Robot – I, Criminal— When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offences", Syracuse Sci. & Tech. L. Rep., 2010; "Virtual Criminal Responsibility", 2011 (Available at SSRN: https://ssrn.com/abstract=1835362 or http://dx.doi.org/10.2139/ssrn.1835362); "Unmanned Vehicles. Subordination to Criminal Law Under the Modern Concept of Criminal Liability", *Journal of Law, Information and Science,* Vol 21, No. 2002, 2012; *When Robots Kill. Artificial Intelligence under Criminal Law*,

earliest work on the topic dates back to 2010) and is the first who attempted at setting out how modern legal frameworks could be applied to AI.[208]

According to Hallevy, it is possible to identify a new kind of offender, i.e., the delinquent thinking machine. It is not relevant for criminal law, he argues, that an offender possesses every human skill. It does not have to be an ideal thinking machine, such as humans. Notably, it is only required that a match between criminal law requirements and the relevant skills and abilities of the system exists.[209]

In his earlier work, Hallevy drew heavily on the comparison between AI systems and corporations, claiming that there is "no substantive legal difference between the idea of criminal liability imposed on corporations and on AI entities".[210] In his latest work, i.e., the book "Liability for Crimes involving Artificial Intelligence Systems", Hallevy detached from his previous publications on different levels, adopting at times a more moderate approach. In this monograph, the author first deconstructs the element of modern criminal liability (external or *actus reus* element and internal or *mens rea* element), and then discusses whether AI technology can fulfill the requirements of each element.

With regards to the element of *actus reus*, Hallevy, as Lagioia and Sartor,[211] without any second thought abandons the doctrine which submits that the criminal act need be the result of a willed bodily movement. As a matter of fact, he considers it a "mongrel requirement" belonging to the past.[212] In other words, according to this author there is no space for any mental element requirement when examining *actus reus*.[213] Notwithstanding of whether one

---

Northeastern University Press, 2013; *Liability for Crimes Involving Artificial Intelligence Systems*, Springer, 2015; "Dangerous Robots – Artificial Intelligence vs. Human Intelligence", 2018 (Available at SSRN: https://ssrn.com/abstract=3121905); "The Basic Models of Criminal Liability of AI Systems and Outer Circles", 2019 (Available at SSRN: https://ssrn.com/abstract=3402527 or http://dx.doi.org/10.2139/ssrn.3402527). Most of the discussion in this research will focus on the 2015 book, which, in the very words of the author, is meant as a "full academic generalization" of the issue (see v., Preface).

[208] R. Charney, "Can Androids Plead Automatism - A Review of When Robots Kill: Artificial Intelligence under the Criminal Law by Gabriel Hallevy", *U. Toronto Fac. L. Rev.*, Vol. 73 , 2015, p. 69.

[209] Hallevy, 2015, p. 25.

[210] Hallevy, "The Criminal Liability of Artificial Intelligence Entities - From Science Fiction to Legal Social Control, 2010, p. 201. The analogy will be dealt with in depth in Ch. 6.4.

[211] See Para. 3.3.2.

[212] Hallevy, 2015, p. 61.

[213] This is far from being the dominant theory, at least in Western criminal law. See *ex multis* Keiler, who states that "Conduct can only be an expression of human agency if it is linked to human will, and

agrees or not with Hallevy's position, one can observe that theories based on willful conduct, since they require a link between "mind and body",[214] represent for a fact the cultural heritage of a reflection which has been based for decades on human or, in the case of corporations, on proxies for humans. One could ask herself whether AI systems indeed warrant for a detachment from this line of reasoning because they are not human, nor proxies for humans. This detachment could take form, as argued by Hallevy, through discarding any sort of mental element when looking at the *actus reus*. Or else, it could lead to theorizing a new form of willful conduct which reflects the functioning of modern AI systems, similarly to what is hypothesized by Lagioia and Sartor or by Abbott and Sarch concerning intent.[215] This reformulation operation of the concept of "willful conduct" is, apropos, what has been done by those criminal justice systems that allow for the liability of corporations.[216]

Following this reasoning, Hallevy claims that "for the question of performing an act in order to satisfy the conduct component requirement, any material performance through factual–external presentation is considered an act, whether the physical performer is strong artificial intelligence entity or not".[217] The same reasoning is also applied to commission-by-omission scenarios. From this statement, it follows that even the simplest machines could perform "conduct under the definition and requirements of criminal".[218]

Concerning the causal nexus between the conduct and the harmful results, Hallevy adheres to a *conditio sine qua non* theory of causation.[219] Consequently, he claims that since AI technology is "capable of committing conduct of all kinds, in the context of criminal law, it is capable of causing results out of this conduct".[220]

---

consciousness i.e., control. […] it is not so much movement that matters, but rather conduct that lies within one's control and reflects the person as a rational agent. If such a link is absent […] the imposition of criminal liability is unwarranted. […] these situations are not merely a denial of mens rea, but more profound, namely the denial of a criminally relevant conduct". J. Keiler, *Actus reus and participation in European Criminal Law,* Intersentia, 2013, p. 61.

214 H.L.A. Hart, *Punishment and Responsibility*, 2nd Ed., Oxford University Press, 2008, p. 97.

215 See Sec. 3.3.1.

216 See Sec. 6.4.

217 Hallevy, 2015, p. 62.

218 Ivi, p. 63.

219 Also referred to as "ultimate cause theory" or "but for" test. According to this theory, to prove that the conduct A cause the result B, the judge has to remove element A from the factors which led to the event and ask herself if event B would still have happened.

220 Hallevy, 2015, p. 66.

When discussing *mens rea*, the author asks himself whether AI technology has, on one hand, the capability of "being aware of conduct, circumstances or possibility of the results' occurrence, in the context of criminal law"[221] and, on the other hand, the capability of consolidating will.[222] In other words, he questions whether AI systems could fulfill the cognition and volitional element of intent. As already mentioned, the author answers positively to both questions. This is an instance of what Pagallo and Barfield referred to as an "à la Hallevy's position"[223], i.e., the humanization of AI-systems.

With regards to ascertaining awareness, he divides the process in two steps. The first step comprises of "absorbing the factual data by senses".[224] The second step entails being able to create a general picture from the information and to use it.

According to Hallevy, AI systems are capable of fulfilling the awareness requirement much better than humans. To prove this, he provides an interesting example based on the differences between a human and an AI-based guard. An AI-based robot guard can scan its surroundings to identify suspicious behavior. When the AI systems detects a movement or a sound, it poses a question to identify the figure it encountered ("Who is there? Please identify yourself!") and then verifies the answer by recurring to its memory. Same as a prison guard would ask the intruder to identify herself and then analyze the answer (the voice and the image of the possible intruder, together with the information provided) to see if it matches individuals in her memory. The most striking difference between a human and a robot guard, though, is in their data collection and analysis potential. Simply put, an AI does not fall asleep and is therefore in an eternal state of high awareness.[225] For these reasons, Hallevy contends that AI systems are able to pass the "awareness-test" with maximum grades.

With regards to proving volition, the author, as Lagioia and Sartor,[226] focuses the discussion on the so-called foreseeability rule presumption. The presumption can be summarized as follows: intent in human offenders is proven if the offender "during the aware commission of the conduct, has foreseen the occurrence of the results as a very high

---

221 Ivi, p. 89.
222 Hallevy, 2015, p. 94.
223 Pagallo & Barfield, 2020, p. 140.
224 Hallevy, 2015, p. 89.
225 Ivi, p. 90.
226 See Para. 3.3.2.

probability option".[227] He contends that resorting to this type of "evidential substitute" is mandated by the fact that "aware will" relates to future situations, where instead awareness relates to facts. In other words: "[a]wareness is rational and realistic, whereas intent is not necessarily. For instance, a person may intend to become an elephant, but that person cannot be aware of being an elephant, as he is not one".[228] It follows that since AI systems can assess that the probabilities that one event (be it winning a chess game or killing a human being) will result from their conduct are very high, and since they can choose to act in pursuant to this assessment, they fulfill the conditions for the foreseeability rule presumption. Hence, it can be presumed that they have intent.[229] Following a parallel logic, Hallevy concludes that AI systems are also capable of forming negligence.[230]

The author then advanced three liability models, which could be used to impose criminal liability on AI systems:

(1) the direct liability model;
(2) the perpetration-through-another model;
(3) the natural probable consequence model.

We can then identify three corresponding scenarios in Hallevy's work which describe the scope of application of the three models:

---

[227] Hallevy, 2015, p. 96.

[228] Ibid.

[229] Hallevy, 2015, p. 98. Hallevy theorizes that it is possible to ascertain an AI system's negligence by analyzing the machine learning process which led to the mistaken decision. He defines negligence as "unawareness of the factual component in spite of the capability to form awareness, when reasonable person could and should have been aware of that component" (p. 124). He summarizes the test in three general questions, which would have to been ascertained by a court (with the aid of a computer scientist expert):

(a) Was the artificial intelligence system unaware of the factual component?

(b) Has the artificial intelligence system the general capability of consolidating awareness of the factual component?

(c) Could a reasonable person have been aware of the factual component?

[230] Ivi, pp. 120-131.

(a) The AI system makes a decision to commit an offense based on its own accumulated experience or knowledge or based on advanced calculations of probabilities – liability model (1);
(b) The AI system is used by a human being as a (sophisticated) tool to commit an offense – liability model (2);
(c) The AI system was not designed to commit the specific offense, but the offense was committed by the artificial intelligence technology nonetheless – liability model (3).[231]

The models can be applied separately or in a combined manner.[232] The direct liability model (1) is applicable in situations similar to the one described at (a). This model is the result of an operation of "humanization of AI systems" which was outlined above. Indeed, Hallevy tested whether the requirements of criminal law could be fulfilled by AI technologies, and concluded the test with positive answers. By doing so, he opened the "gate for imposition of criminal liability upon artificial intelligence technology as direct offenders".[233] This model confers to Hallevy the title of AI punishment expansionist.

The perpetration-through-another model (2) is suitable to impose criminal liability on a human offender (identified by Hallevy in the programmer or the user) in situations such as the one described at (b). The AI system is treated as an innocent agent and therefore the model considers the action committed materially by the AI system as if it had been the action of the user or of the programmer.[234] Accordingly, this model is not suitable for application in situations such as (a) and (c).

The natural probable consequence model (3) kicks in situations such as the one described at (c), i.e., when there is a "deep involvement of the programmers or users in the AI entity's daily activities, but without any intention of committing any offense via the AI

[231] Hallevy, 2015, p. 112.
[232] In his book, Hallevy divides the argument differently, analyzing the different models of liability when discussing in detail the mental health requirements. For the sake of this analysis, we will discuss these models at once, following the structure adopted in the earlier paper "The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control", 2010.
[233] Hallevy, 2015, p. 105.
[234] Hallevy, "The Criminal Liability of Artificial Intelligence Entities - From Science Fiction to Legal Social Control, 2010, p. 180.

entity".[235] It is necessary at this point to make a further distinction to fully grasp the rationale behind Hallevy's theory. We will refer to three sub-situations:

(c)(i). the programmer had no criminal intent whatsoever and she was not negligent in the programming of the AI system;

(c)(ii). the programmer had no criminal intent whatsoever, but she was negligent in the programming of the AI system;

(c)(iii). the programmer designed the AI system to commit an offense, but the system diverged from the plan causing either more or different harm than the predicted offense (*aberratio delicti*);

In situation (c)(i) the programmer was not negligent, hence it would be inappropriate to apply this model and in general to enforce any kind of criminal response upon her. Conversely, situation (c)(ii) is a case of "pure negligence" and therefore does not fall in the application scope of model (3), but should follow the standards set for negligence crimes.[236] Finally, situation (c)(iii) represents, according to Hallevy, the ideal scenario for the natural probable consequence model, since this model "is meant to deal with unplanned developments of a planned delinquent event".[237] In cases like these, the programmer or the user shall be deemed responsible for the crime which was committed by the AI system in addition or in substitution of the planned crime. Hallevy provides the following example:

> [A] medical expert artificial intelligence system is used for diagnosis of certain types of diseases through analyzing the patient's symptoms. The artificial intelligence system analysis is based on machine learning, which inductively analyses and generalizes specific cases. The system fails to diagnose correctly one case, and that reveals to wrong treatment, which worsens the patient's situation and finally causes the patient's death. The analysis of the artificial intelligence system's activity reveals negligence of it, and it fulfills both factual and mental elements requirements of the relevant negligence offense (negligent homicide). At this point arises the question of the *programmer*'s

[235] Ivi, p. 181.
[236] Hallevy, "The Criminal Liability of Artificial Intelligence Entities - From Science Fiction to Legal Social Control, 2010, p. 184.
[237] Hallevy, 2015, p. 119.

> criminal liability for that offense. His criminal liability is not related to the decision to use the artificial intelligence system, to follow its diagnosis, etc., but it is related to the very initial programming of the system. If the programmer would have programmed the system to kill patients and instrumentally used it for this purpose, it would have been perpetration-through-another of murder, but this is not the case here. For the programmer's criminal liability in this case the probable consequence liability may be relevant.[238]

What about combining the different liability models? In cases such as the situation described at (c)(iii), assuming that the AI system was not used as an innocent agent, the direct liability model could be applied to attribute criminal liability upon the AI system in addition to the one attributable to human operators.

Another example of a possible combination is the case where the programmer of the AI system is itself an AI system. Imagine that the system X programs a system Y to illegally access a computer, and that Y fulfills this criminal task, but then intentionally uses the illegal access to steal the victim's identity to commit a series of online frauds. In this case, X would be liable for Y's behavior following the natural probable consequence model, where instead Y would be liable for its own behavior according to the direct liability model.

According to some, Hallevy fails to recognize that "while certain acts can be attributed to the heads of each company, attributing each line of code and task to individual programmers is a monumental task".[239] Moreover, he neglects "the criminal liability of hardware manufacturers",[240] who would not be liable under any of the three models.

Finally, the author focuses on punishment.[241] On the one hand, he argues that retribution and deterrence would prove useless in the case of punishing robots but could prove valuable when punishing the human participants in the offence.[242] On the other hand,

---

[238] Hallevy, 2015, p. 134.

[239] D.J. J. Kim, "Artificial intelligence and crime: what killer robots could teach about criminal law", Thesis in Law, Victoria University of Wellington, 2017, p. 26. Available at: https://researcharchive.vuw.ac.nz/xmlui/bitstream/handle/10063/7927/paper_access.pdf?sequence=1

[240] Ibid.

[241] Hallevy also outlines how specific punishments (capital penalty, imprisonment and suspended imprisonment, probation, public service, fine) would be applied upon AI systems (Ch. 6.2.2.).

[242] Hallevy, 2015, p. 210.

he states that rehabilitation and incapacitation are relevant from an AI-punishment perspective. Rehabilitation could function for machines as it functions for humans: it may be used to refine the machine learning process, as a way to lead AI systems to make better decisions. The rehabilitated AI system, then, would be able to perform better, same as rehabilitated defendants should have better tools to face reality.[243] Incapacitation would be the last resort measure directed at those systems that have proven to be incapable of changing their ways through their inner processes (i.e., via machine learning).[244]

3.2.2 Ying Hu: a Criminal Code for Robots

Hu makes a positive case for imposing direct criminal liability on "smart robots", that is, robots that fulfil three threshold conditions. These conditions are: "the robot must be (1) equipped with algorithms that can make nontrivial morally relevant decisions; (2) capable of communicating its moral decisions to humans; and (3) permitted to act on its environment without immediate human supervision".[245]

Hu shares one research question with Abbott: are there any good reasons which justify imposing criminal liability on (smart) robots? Abbott believes that it would be possible to build a coherent theoretical case for punishing AI but that it is simply a bad idea in light of the existence of less disruptive alternatives. [246] Similarly, Hu argues that there can be good reasons which justify imposing criminal liability on smart robots. Her conclusions, though, are different from Abbott's. She proposes the introduction of a new criminal code for robots.

Indeed, Hu fits wholly in the Front of Robotic Liberation, as she argues in favor of legal personhood for robots. Yet, she distances herself strongly from its paramount representative, i.e., Hallevy. The Israeli author, she claims, fails at describing in detail the type of robot on which to impose criminal liability and at explaining why we should adopt the criminal tool in the first place.[247] Moreover, he "appears to assume that, since we already impose criminal liability on non-human entities such as corporations, extending such liability to robots requires little justification".[248] Furthermore, Hu also addresses two representatives

[243] Ivi, p. 211.
[244] Hallevy, 2015, p. 211.
[245] Hu, 2019, p. 490.
[246] See Section 3.3.1.
[247] Hu, 2019, p. 492, note 13, referring to Hallevy, 2013, pp. 38 and 66.
[248] Hu, 2019, p. 492.

of the skeptic front, i.e., Gless, Silverman and Weigend,[249] complaining that they only restrict their analysis to existing robots.[250]

The author acknowledges beforehand that her analysis might be speculative, yet she believes that its value lies in the fact that it is an informative tool, which could prove useful in the regulation vs. innovation race.[251] Moreover, it could work as tenet for scientists working on moral machines, when it comes to deciding which moral norms should be taught to the robots and which training data they should adopt.

Let now us explain the aforementioned conditions for labelling a robot as "smart". The first condition regards the creation of "moral algorithms" capable of making "nontrivial morally relevant decisions" (meaning decisions "between or among two or more courses of actions that might be considered right or wrong by ordinary members of our society").[252] Indeed, as the author rightly recognizes, moral decision have acquired importance especially in the field of autonomous driving.[253] Moral machines can be built either following a rule-based approach, which requires programmers to encode in the algorithm all the moral rules *ex ante*,[254] or a utility-maximization approach, where machines learn moral rules by trial and error through reinforcement learning. Regardless of the approach used to create moral machines, the second condition implies that the robot communicates these decisions to humans, together with the alternative courses of actions that were available, and the weight given by robot to each. The third condition necessitates that the human-in-the-loop partakes

---

[249] See Sec. 3.4.5.1.

[250] Hu, 2019, p. 492, note 13, referring to S. Gless, E. Silverman & T. Weigend, "If Robots Cause Harm, who Is to Blame: Self-Driving Cars and Criminal Liability", *New Criminal Law Review*, Vol. 19, No. 3, 2016, p. 423.

[251] "Although our analysis might be speculative in some respects, the thought experiment is nevertheless invaluable: It helps identify the key considerations that should inform our decision whether to impose criminal liability on robots. We will, in turn, be better positioned to decide whether and when to apply criminal liability to robots, as technological advances push us closer to the turning point in that spectrum", Hu, 2019, p. 495.

[252] Hu, 2019, p. 496.

[253] Think for example of the MIT Moral Machine. See Para. 5.3.2.1.

[254] The author then distinguishes between strict rule-based approaches, which do not allow the robot to learn new moral rules or to make decisions when there is "ethical uncertainty"; and soft rule-based approaches, where the robot is equipped with a set of "high-level moral rules" and is also fed examples which demonstrate how to apply these rules to real cases. The purpose of the examples is for the robot to learn new principles which can be applied to new scenarios. Hu, 2019, p. 497.

a distant position from the smart robot. In other words, there should be no "immediate human supervision" on the action of the smart robot.[255]

Having established this, Hu turns to the heart of her argument: the creation of a Criminal Code for Robots. The reasons for its introduction are twofold: first, there are grounds to hold smart robots to a higher moral standard than humans. To support her claim, Hu argues that smart robots could be held liable for failure to act not only when they have a legal duty to do so.[256] Yet, this would entail diverting from one of criminal law's founding principles, that is, the principle of culpability. Second, smart robots might prompt new moral questions which were never faced by humans, as they are able to act in ways that are physically impossible for a human being.

What would be the benefits, then, of creating a Criminal Code for Robots? Hu claims that it would introduce a minimum set of moral standards decided collectively by society. What is more, it would work as a legal basis for holding "robot manufacturers" (subjects who participate in creating the algorithms) and "robot trainers" (subjects who train the algorithms) criminally liable for failing the duty of care, i.e., for preventing smart robots from behaving in a way that is against the code.

One counter argument which prompts instinctively at this point is: if one of the reasons to introduce criminal norms for robots is that they are capable of behavior that is not even conceivable by humans, how could the very same humans be punished for these behaviors? As a matter of fact, Hu does not explain how these new moral issues should be dealt with.

She subsequently presents her case for imposing criminal liability on robots based on three arguments. Firstly, criminal punishment has a censuring (or expressive) function. It communicates the disapproval of the community towards a morally wrongful conduct and this function acquires greater importance when no human being is at fault for the robot's misconduct. [257] Secondly, punishing robots would provide emotional relief to victims of smart robots' misbehavior. Thirdly, it would be of use to identify culpable (human) individuals, since those who are not at fault would be inclined to cooperate with investigations, therefore pinpointing to those who are. It would also nudge towards the

[255] Hu, 2019, p. 499.
[256] Think of the example discussed at 1.7, D.
[257] Hu, 2019, p. 490.

creation of self-policing mechanisms for robot manufactures and users, who would put in place safeguards against robot harm to avoid sanctions on the robot.

Hu, then, addresses five possible objections to her arguments. First: robot criminal liability is redundant in respect to imposing criminal liability on individuals responsible for robot misconduct. Hu claims that this is not the case, since if we were to hold the latter liable it would be for an omission, i.e., it would be for failing negligently or recklessly to prevent the robot from harmful conduct, rather than for the harmful conduct per se. Furthermore, it would not be redundant in all those cases where neither the smart robot manufacturers nor the trainers are at fault, same as "situations in which *[in American jurisprudence]* corporations are held liable for offenses despite the fact that none of their human agents are held liable of those offences".[258]

Second: robots are incapable of performing actions since they lack the capacity to form any mental state. She responds by relying on Peter French's account of corporate liability: one can claim that *x* did *y* intentionally if *x* had a reason for doing *y,* which was the cause of doing it. This entails that corporations are able to act intentionally, since they have an interest in performing an action *y* if it is likely to result in the realization of the corporate goals and policies.[259] In order for the act to be defined as intentionally pursued by the corporation, it has to be done for corporate reasons, i.e., to pursue a corporate policy established by an internal decision structure. Accordingly, smart robots possess moral goals and policies which represent the views of robot trainers (similar to the internal decision structure of a corporation). Hence, whenever they act, they do so follow their moral algorithms, that is, they act intentionally.

Addressing this objection might be relevant to respond to those who believe that acting "willingly" entails that the agent must be capable of forming intent. However, Hu does not address a big share of the debate represented by those who believe that the "vital link between mind and body",[260] i.e., the rational capacities which are necessary to consider a conduct voluntary and therefore punishable, has little or nothing to do with *mens rea*. Indeed, she claims that the "standard" conception of action is based on intentionality, which is a mental state and, as a consequence, one cannot be the agent of an action if she lacks the

---

[258] Hu, 2019, p. 515.
[259] French P.A., *Collective and Corporate Responsibility*, Columbia University Press, 1984, p. 40.
[260] Hart, 2008, p. 107.

capacity to form mental states.[261] But who defines the standard concept of a criminally relevant act? The question remains open.

Third: robots are incapable of performing morally wrongful actions, since they are not capable to understand that their actions were wrong. Hu's response can be summarized as follows: since smart robots are led by moral algorithms, trained by individuals who know the moral values of our society, they are members of our moral community. This entails, then, that if they commit a moral wrong, they do so intentionally.

Fourth: whether it might be true that smart robots can behave morally, they do not do so autonomously, since the principles they act upon were engineered by its manufacturers and trainers. Hu argues that those who support this objection adhere to Kant's conception of a moral agent, i.e., an agent must be autonomous in the sense that it must be the author of his desires. She maintains, then, that we should look at alternative theories of moral agency, such as the one of collective responsibility as elaborated by List and Pettit. According to this theory an agent can be held responsible of her misbehavior in case three conditions are fulfilled: one, that the agent faces a "normatively significant choice" which entails picking between good/bad, right/wrong, etc.; two, that the agent is capable to understand and access information on how to make said choice; three, that the agent "has the control required for choosing between options".[262] List and Pettit argue that these conditions can be satisfied by groups, such as corporations. Hu, then, maintains that smart robots can fulfill these requirements as well. Specifically, regarding the third one, she argues that smart robots have a certain kind of autonomy on making a normatively significant choice. As a matter of fact, they are capable of applying the moral norms which were taught to them by the manufactures and the trainers to new situations.

Fifth: recognizing robots as legal persons would be harmful, as it would encourage people to anthropomorphize robots. Hu discards this argument quite rapidly by affirming that the concerns on antrophormizing robots are not strong enough to outweigh the positive benefits of holding smart robots criminally liable.

In conclusion, Hu focuses on AI-punishment. The author, as well as Abbott,[263] adopts H.L.A. Hart's definition of punishment:

---

[261] Hu, 2019, p. 518.
[262] C. List & P. Petit, *Group Agency*, Oxford University Press, 2011, p. 155.
[263] R. Abbott, *The Reasonable Robot*, Cambridge University Press, 2020, p. 115.

(1) it must involve pain or other consequences normally considered unpleasant;
(2) it must be for an offense against legal rules;
(3) it must be of an actual or supposed offender for his offense;
(4) it must be intentionally administered by humans other than the offender; and
(5) it must be imposed and administered by an authority constituted by a legal system against which the offense is committed.[264]

She argues that conditions (2), (3), (4) and (5) can be easily fulfilled. With regards to (1), she contends that punishment should be deemed as unpleasant by the general members of our community. Interestingly so, Hu is amongst the few legal scholars who theorizes how criminal punishment towards AI systems could take from in practice. She hypothesizes four punishments:

(1) physically destroying the robot (the robot equivalent of a "death sentence");
(2) destroying or re-writing the moral algorithms of the robot (the robot equivalent of a "hospital order");
(3) preventing the robot from being put to use (the robot equivalent of a "prison sentence"); and/or
(4) ordering fines to be paid out of the insurance fund (the robot equivalent of a "fine").[265]

3.2.3 Christina Mulligan: Revenge Against Robots

In her paper "Revenge Against Robots" Mulligan argues that imposing punishment ("revenge" or "vengeance", as defined by the author) on AI systems would result in retributive benefits consisting in the psychological satisfaction of victims of AI-misbehavior.[266] Whether her considerations are not specifically tailored to criminal sanctions,

[264] Hart, 2008, pp. 4-5.
[265] Hu, 2019, p. 529.
[266] C. Mulligan, "Revenge Against Robots", *South Carolina Law Review*, Vol. 69, 2018, p. 578.

some of her reflections are indeed valuable to the criminal legal debate, and this is proven by the fact that certain scholars mentioned in this Chapter refer to her work.[267]

According to Mulligan, in the future it will be impossible to say that the "rogue" behavior of robots based on black box algorithms[268] is "caused" by the manufacturer, the programmer or the seller. These opaque algorithms occasion the impossibility to explain why robots behaved in a certain way, even through an "autopsy" [269] of the robot's algorithm performed by the most skilled AI-coroner. It might be possible to affirm that the actions of programmers and manufactures are but-for causes (i.e., the *condiciones sine quibus non*) for the realization of the harmful event, yet they cannot be seen as proximate causes, nor as reasonably foreseeable ones. The unpredicted robot behavior which led to the (criminal) outcome should be seen as triggered by intervening causes, that is, experiences that affected its learning mechanisms. According to Mulligan it follows, then, that the robot represents the proximate cause of the unwanted event, as "any other answer would torture the meaning of 'proximate cause' ".[270]

Having established this, Mulligan moves on to analyze how the question of moral culpability, i.e., if robots are capable of moral behavior, affects her argument in favor of robot punishment. She analyzes two questions: first, "does the sense that blameworthiness supervenes on the existence of 'free will' change whether robots should be punished for their actions to satisfy their victims?"; second, "does it change whether the victim *feels* that the punishment is morally justified?". The answer is negative to both questions. She states concisely:

> Either a robot is as morally blameworthy and as deserving of penalty or other legal action as a human, or the robot is like a rock and is neither deserving nor undeserving

---

[267] Abbott, 2020, p. 117; Hu, 2019, p. 531.

[268] See also the definition of W. Nicholson Price II, "Big Data, Patents, and the Future of Medicine", *Cardozo Law Review*, Vol. 37, 2016, p. 1404 who describes box algorithms (that analyze health information) as "'black-box' precisely because the relationships at [*their*] heart are opaque not because their developers deliberately hide them, but because either they are too complex to understand, or they are the product of non-transparent algorithms that never tell scientists, 'this is what we found'. Opacity is not desirable but is rather a necessary byproduct of the development process". See also Sec. 6.3.2.2 for an analysis of the black box phenomenon.

[269] Mulligan, 2018, p. 590.

[270] Ibid.

> of any sort of treatment. In both situations, the robot's moral status does not supply a reason to avoid taking action against it, given the presence of other reasons to do so.[271]

Lastly, Mulligan focuses on punishment. Asaro[272] is off track, she states, in claiming that punishing robots would not achieve the classical goal of punishment such as retribution, reform or deterrence. Indeed, the goal of criminal law is to create psychological satisfaction for the victims of the robot, and this can be done most effectively by introducing a modern version of the Middle Age practice of "noxal surrender". Noxal surrender would involve handing over the faulty robot to the victim or to her family so that they to do what they think its best with it in order to fit their satisfaction. As a matter of fact, a wronged party "may indeed be quite justified in dragging a robot out into an empty field and walloping it with a baseball bat".[273] All things considered, Mulligan presents, as she defines it herself, quite "an outlandish argument".[274]

3.2.4 Lasse Quarck: the German Exception

Quarck represents quite the exception: he is the only German author quoted in this Chapter who is part of the expansionist front.[275] He believes that it is unacceptable for criminal law to remain behind during the process of digital transformation that we are living. Consequently, he argues that introducing AI criminal liability will be unavoidable in the long run. Scholars must discuss right now the direct criminal liability of the intelligent agent, regardless of whether they think that human-like AI systems will come to life or not soon.

Quarck focuses on three dogmatic challenges regarding the application of criminal liability to AI agents: first, the *actus reus* challenge; second, the *mens rea* challenge: third, the punishability challenge.

---

[271] Mulligan, 2018, p. 593.
[272] See Sec. 3.4.4.
[273] Mulligan, 2018, p. 595.
[274] Ibid. For an interesting analysis on the relationship between people's view on moral judgments and punishment of automated systems vs existing legal doctrines see the empirical study conducted G. Lima et al., "The Conflict Between People's Urge to Punish AI and Legal Systems", *Front. Robot. AI*, Vol. 8, 2021. The study will be discussed at Para.6.5.
[275] Quarck, 2020.

"*Das Strafrecht ist von Menschen für Menschen erdacht worden"*[276] claims Quarck. It follows that the conceptual categories of criminal law that we know of cannot be transferred directly to AI agents. With regards to the *actus reus* challenge, he argues that what is relevant is how strongly one would like to "normatively charge"[277] the concept of agency: if one believes that the ability of the AI to understand norms is a prerequisite for its ability to act, then we would have to refute the idea that current AI systems could fulfil the *actus reus* requirement (which is the argument made by Gerard Seher, exponent of the skeptics front, in a nutshell).[278] He objects to Seher by claiming that indeed the requirement of "willful" conduct can be dispensed with if we base criminal liability on systemic (or algorithmic) wrongdoing rather than on the commission of a singular individual, similarly to what is done in those legal systems which are acquainted with corporate criminal liability.[279]

With regards to the *mens rea* challenge, Quarck believes that free will does not represent a mandatory prerequisite to establish the culpability of an agent.[280] Indeed, he argues, free will is only attributed since it cannot be proven that it actually exists (or that it doesn't). Guilt is assigned to an agent to resolve a conflict that has been caused in our society by the commission of the wrongdoing, because no harm can go without a sanction. Following such a consequentialist approach, Quarck claims that the same guilt-assigning operation can be conducted with algorithmic misconduct. It does not matter whether intelligent or human agents commit a crime because of a biological or algorithmic process, or because of a dysfunction in the formation of their free will: it is our social system that decides to whom we must assign criminal responsibility.[281]

With regards to the punishability challenge, if an AI system can fulfill both *mens rea* and *actus reus* criteria, then punishing AI systems would fulfil both general and special prevention purposes. With regards to the former, the AI system could be reprogrammed to include the meaning of the standard violated through the criminal conduct. With regards to the latter, assuming that the punished AI systems were in a network with other agents, this would cause

---

[276] Ivi, p. 57.

[277] "*Es könnte zunächst darauf ankommen, inwieweit man den Begriff normativ auflädt. Sieht man die Fähigkeit der KI zum wenigstens potenziellen Normverständnis als Voraussetzung der Handlungsfähigkeit, so wäre diese, zumindest zum jetzigen Zeitpunkt, abzulehnen*". Quarck, 2020, p. 66.

[278] See Para. 3.4.5.3.

[279] Quarck, 2020, p. 65.

[280] Ivi, p. 68.

[281] Quarck, 2020, p. 68.

the other networked systems to include the same implementation (*rectius*, correction) in their own code.

3.3 MODERATES

On the one hand, the authors placed in this category recognize that there is a mismatch between the law, which is designed to regulate human behavior, and the ways algorithmic behavior takes place. Hence, they advocate for a change in the law (or in the interpretation of the law).[282] On the other, these authors only suggest a moderate change and therefore do not disrupt completely the traditional foundations of criminal law. A further characteristic shared by some of the authors which are mentioned in this section (Abbott & Sarch, Lagioia & Sartor), is that their theories are the product of hybridization, i.e., a mixture between concept of criminal law and concepts stemming from other disciplines, such as ethics, psychology, and computer science.

Indeed, much of the academic literature analyzed in this Chapter is the byproduct of an old discussion, which can be reconducted to legal moralism.[283] What comes first: moral values or law?[284] It is tempting, at times, to confuse rules of law and rules of morality, especially when it comes to criminal law. The topic has been addressed by conspicuous literature over the past decades.[285] As a consequence, the aforementioned classical debate found renewed importance with the incurrence of the discussion on how to regulate AI. It is also in this light that the authors analyzed here acquire value.

3.3.1 Ryan Abbott and Alex Sarch: a General Theory for AI-Punishment

Two of the most prominent authors which take a moderate stance on AI and criminal liability are certainly Abbott and Sarch with their paper "Punishing Artificial Intelligence: Legal

[282] Abbott, 2020, p. 3.
[283] See Ch. 4.2. for a discussion of theories of criminalization.
[284] See S. Rodotà, "Etica e Diritto (dialogo tra alcuni studenti e Stefano Rodotà) con una Presentazione di Gaetano Azzariti", *Costituzionalismo.it,* Vol. 1, 2019.
[285] J. Allan, "Revisiting the Hart-Devlin Debate: At the Periphery and By the Numbers", *San Diego L. Rev.*, Vol. 54, 2017, p. 423.

Fiction or Science Fiction".[286] In this paper, they focus on justifications for criminal punishment, rather than on the specific elements of criminal liability, and they adopt a very pragmatic approach. Specifically, they reflect on whether the doctrinal and theoretical commitments of criminal law can be reconciled with criminal liability for AI. Notably, the authors do not rule out direct punishment of AI systems completely.

The reasons why the work of Abbott and Sarch is relevant are manifold. To begin with, they were the first ones to introduce to the (criminal law) academic discourse the famous case of the Random Darknet Shopper (RDS),[287] later used by several authors.[288]

The RDS was an online bot initially installed in a nonprofit gallery in St. Gallen, Switzerland, which was created to act as an (a)live piece of an art exhibition.[289] The software was programmed to autonomously spend 100$ in bitcoin per week on a deep web market named Agora Shop. Each week the bot randomly chose an object to purchase, which was then sent directly to the gallery. The nature of the object would be disclosed only once the package had arrived at the exhibition space, where it was unpacked and then put on display. From October 2014 to January 2015 the RDS bought a total of 12 items. Outstandingly, some of them included counterfeit Louis Vuitton Handbag, counterfeit Nike Air Yeezy shoes, a fake Hungarian passport and twelve ecstasy pills. In January 2015, the Swiss police

---

[286] Abbott & Sarch, 2019. Ryan Abbott expanded his reflections in his subsequent publication, *The Reasonable Robot,* 2020

[287] J. Lackman, Random Darknet Shopper, *Aksioma – Institute for Contemporary Art,* 2016. Available at: https://aksioma.org/pdf/aksioma_PostScriptUM_23_ENG_Bitnik.pdf. See also: !Mediengruppe Bitnik, Random Darknet Shopper, 2014-2016. Available at: https://wwwwwwwwwwwwwwwwwwwww.bitnik.org/r/; M. Power, "What happens when a software bot goes on a Darknet shopping spree?", *The Guardian*, 5 December 2014. Available at: www.theguardian.com/technology/2014/dec/05/software-bot-darknet-shopping-spree-random-shopper; K. Grant, "Random darknet shopper exhibition featuring automated dark web purchases opens in London", *The Independent*, 2 December 2015. Available at: https://www.independent.co.uk/life-style/gadgets-and-tech/news/random-darknet-shopper-exhibition-featuring-automated-dark-web-purchases-opens-in-london-a6770316.html.

[288] F. Lagioia & G. Sartor, "AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective", *Philosophy & Technology*, Vol. 33, 2020; Hayward & Maas, 2020; D. J. Baker & P. H. Robinson, *Artificial intelligence and the Law. Cybercrime and Criminal Liability*, Routledge, 2021; J. Turner, *Robot Rules. Regulating Artificial Intelligence*, Gildan Media Corporation, 2019. Abbott further developed his reflections in Ch. 6 of his book "The Reasonable Robot. Artificial Intelligence and the Law", 2020. We will refer to this later work only when it adds to what is stated in his collaborative work with Sarch.

[289] !Mediengruppe Bitnik & !Digital Brainstoarming, "The Darknet – From Memes to Onionland. An Exploration", Kunst Halle Sankt Gallen.

seized RDS and its possessions. Soon later, the charges against the artist and the bot were withdrawn.

To continue, Abbott and Sarch coined the definition of "Hard AI Crime".[290] Before focusing on this definition, it is necessary to take a step back. According to the authors, AI system behavior expresses four main features, which are relevant features from a criminal law perspective:

1) Unpredictability, i.e., the capacity for the system to engage in activities which were not intended or foreseen by its creators;
2) Unexplainability, i.e., the incapacity to explain why the AI system chose a certain pattern of behavior;
3) Autonomy, i.e., the capacity for the AI systems to act independently of human control and therefore to cause harm without being under the direct control of an individual;
4) Complexity, i.e., the fact that the AI system is the output of the contribution of many individuals over a long period or that its conduct is the result of a training based on huge databases coming from heterogeneous sources.

The combination of these factors might lead to irreducibility i.e., the impossibility to reconnect the crime to a liable person. Based on this, the authors distinguish between "AI Crimes" and "Hard AI Crimes". AI Crimes are defined as "cases in which an AI would be criminally liable if a natural person had performed the same act".[291] Hard AI Crimes are instead defined as "scenarios where crimes are functionally committed by machines and there is no identifiable person who has acted with criminal culpability".[292] If one thinks of this distinction in terms of irreducibility, Hard AI Crimes are instances in which harmful AI conduct is not reducible to a human actor, either for practical reasons (because of the difficulty to identify how individuals singularly contributed to the design of the system) or for trivial reasons (because the human misconduct does not meet the threshold required to

[290] The term AI-Crime (AIC) appears also in T. C. King et al., "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions", *Sci Eng Ethics*, Vol. 26, No.1, 2019.
[291] Abbott & Sarch, 2019, p. 332.
[292] Ivi., p. 328.

activate the criminal sanction). In other words, irreducibility entails that the AI criminal act cannot be traced back to the wrongful act of a person.[293] Indeed, according to the authors Hard-AI Crime seems to make "the strongest case for punishing artificial intelligence".[294]

What is more, the authors provide a comprehensive theory for the foundation of criminal punishment of AI systems. They anchor their reflections building on a theory of punishment, which is, in turn, based on affirmative (pluralist) benefits. They contend, indeed, that punishing AI directly might lead to significant affirmative benefits. First, they argue that it could obtain general deterrence. By doing so, they directly address Peter Asaro's undeterrability argument.[295] Their counter argument can be summarized as follows: whether it might be true that AI as we know it is not a moral agent responsive to punishment, and therefore specific deterrence might be unattainable, punishment of AI systems could lead to unrestricted general deterrence of other subjects, namely of developers, owners or users of AI. Moreover, AI punishment could have some expressive benefits, in the sense that it would convey a "message of official condemnation that could reaffirm the interests, rights, and ultimately the value of the victims of the harmful AI".[296] Yet, the authors recognize that this argument based on "folk morality" might incur in certain unsurmountable objections.[297]

Abbott and Sarch address three main challenges which could be brought upon AI punishment from a retributivist point of view, namely:

(1) The eligibility challenge,
(2) the reducibility challenge, and
(3) the spillover objection.

---

[293] Abbott, 2020, p. 13.
[294] Abbott, 2020, p. 112.
[295] Abbott and Sarch claim that Asaro failed to distinguish between general and special deterrence. Cfr., "And finally, deterrence only makes sense when moral agents are capable of recognizing the similarity of their potential choices and actions to those of other moral agents who have been punished for the wrong choices and actions-without this reflexivity of choice by a moral agent, and recognition of similarity between and among moral agents, punishment cannot possibly result in deterrence", P. M. Asaro, "A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics", in P. Lin, K. Abney & G. A. Bekey, *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, 2012, p. 181.
[296] Abbott & Sarch, 2019, p. 346. See also Mulligan, 2018, p. 580.
[297] These *prima facie* objections are: it would amount to mob justice and could lead to populist demands for justice; it could lead to more undesirable bad human behavior, and.it would carry affirmative costs for the criminal justice system. Abbott & Sarch, 2019, p. 348.

As it will be higlighted, one trait of the authors becomes evident from how they describe these challenges, and the related responses: their non-continental legal background.

Starting with the eligibility challenge (1), it entails that AI systems cannot satisfy the prerequisites of criminal law, hence they cannot be sanctioned.

It can come in both a narrow and a general form. The narrow eligibility challenge suggests that AI systems cannot possess *mens rea*, where instead the general eligibility challenge holds that an AI system cannot be culpable in a broader perspective (i.e., AI systems lack the capacity to weigh reasons and to deliberate in a blameworthy way). The answers to the eligibility challenge considered by the authors are three: *respondeat superior*, strict liability and providing a framework for direct *mens rea* analysis for AI.

Respondeat superior is a legal fiction tool adopted in the field of corporate liability, which builds a bridge between the mental states of the agent of the corporation and the corporation itself. Differently from what the authors state ("imputation principles of this kind are well-understood and legally accepted"),[298] this legal construct frequently incurs in overwhelming objections from continental lawyers. Besides, the authors discard *respondeat superior* as an inadequate response in cases of Hard AI Crime since, differently from corporations, it is not guaranteed that an AI will come with a "ready supply of identifiable human actors whose mental states can be imputed".[299] In other words corporations, unlike AI systems, are composed of people.[300]

Abbott and Sarch then move on to the "familiar route" option of establishing new strict liability offenses for AI crimes, which could seem more plausible in the case of AI systems, as they are not protected by the just desert constraint. Yet, the authors reckon that there might be cases in which AI misbehavior does not fulfill the *actus reus* requirement of a crime. Specifically, they argue that AI systems are not capable of conscious voluntary action: same as "[a] hurricane cannot perform an act but can cause no shortage of harm".[301] Even if we theorized a new strict liability offense providing for a duty of the AI to non-harm humans, these offenses would bear the cost of diluting the meaning of criminal law, hence its

[298] Abbott & Sarch, 2019, p. 351.
[299] Abbott & Sarch, 2019, p. 352.
[300] Abbott, 2020, p. 14.
[301] Ibid.

expressive benefits. These costs have to be weighted when considering strict liability as a solution to the eligibility challenge.

The third response focuses on whether it would be possible to legally construct AI mental states. Could an AI be culpable (*broad* eligibility challenge)? Could an AI possess a specific *mens rea*, as proscribed by the single offense? When answering these questions, the authors introduce reflections which will be later expanded by Lagioia-Sartor and Ashton.[302]

Primarily, they maintain that the preponderant theory regarding culpability is based on whether the individual manifests "insufficient regard for legally protected interest or values".[303] Accordingly, corporations can be accounted as being directly criminally culpable through the "information-gathering, reasoning and decision-making procedures" of their employees.[304] Can we apply the same type of reasoning to AI systems?

Indeed, AI systems can gather information, process it and determine with autonomy how to complete pre-established goals. Imagine if these systems were programmed to follow certain rules, for example not to hurt humans, and they diverge from this rule to reach their goals.[305] Could this be considered as disregard for the norm, therefore as criminally culpable action?

As the authors themselves notice, their reflections are not far from Hu's expansionist point of view.[306] Yet, they modestly contend that their only focus is legal culpability, not moral responsibility, and that therefore the bar, which has to be reached, is lower.

To address the *strict* eligibility challenge, Abbott and Sarch introduce the "Belief Desire Intention" (BDI) model for intent, which was first ideated in the 1980s in the field of cognitive science by Micheal Bratman[307] As it will be revealed, Lagioia and Sartor's analysis relies heavily on Bratman's concept of intent. For this reason, it will be explained further in depth in the next paragraph. For now, it is relevant to mention that this theory entails that an agent intends an outcome when she guides her conduct in the direction of causing that outcome. The concept of "direction of action" requires further clarifications: it implies that the agent will adjust her behavior to make the outcome more likely and that it will monitor

---

[302] See section 3.3.2 and Sec. 6.2.2.
[303] Abbott & Sarch, 2019, p. 355.
[304] Ibid.
[305] See Lagioia and Sartor, Sec. 3.3.2.
[306] See Sec. 3.2.2.
[307] M. Bratman, *Intention, plans, and Practical Reason*, Harvard University Press, 1987.

the surroundings to find ways, which will increase the probability of the outcome. For example, if a person is driving with the intention to harm a pedestrian, if she detects an obstacle between her car and the victim, she will be willing to change her route in order to avoid the obstacle and obtain her goal. The same reasoning could be applied to an AV. In other words, judges would have to ask themselves whether the system was adjusting and "*guiding* its behavior as to make this outcome more likely".[308] In case of a positive answer, it could be argued that the AI system had the purpose of running over the pedestrian. For all these above reasons, the authors defeat challenge (1).

The reducibility challenge (2) implies that it is always possible to identify the culpable human being behind non-human misbehavior. Whether this could be a stronger argument when it comes to corporations, the same cannot be assumed regarding AI systems. These systems behave in a way that is inherently more distant from the humans behind the machine: as indicated above, they are capable of autonomous, unforeseeable, and unpredictable conduct. Moreover, the reducibility challenge can be overcome also from a criminal policy perspective: legislators would have to criminalize "infinitely fine-grains form of misconducts"[309] such as "momentary lapses of attention, the failure to perceive emerging problems that are difficult to notice, tiny bits of carelessness, mistakes in prioritizing time and resources, not being sufficiently critical of groupthink"[310] and more. For these reasons, the authors defeat the reducibility challenge as well.

Finally, the spillover challenge (3) implies that it would be unfair, from a just desert constrain perspective, for the state to punish innocent human bystanders for crimes committed by the AI system. Yet, the authors argue, this is not an issue belonging exclusively to AI or to corporate liability: it is a general issue of criminal law, which has to do with the effectiveness of punishment. In other words, "[i]t is an omnipresent problem with criminal punishment, which should be addressed for any novel mode of criminal punishment - whether for corporations or AI".[311] Based on this, the authors negate the spillover challenge.

To conclude, the authors believe that while it is possible to make a coherent theoretical case for punishing AI, it is not justified considering the existence of less "disruptive"

---

[308] Abbott & Sarch, 2019, p. 358.
[309] Ivi, p. 362.
[310] Abbott & Sarch, 2019, p. 362.
[311] Ivi, p. 364.

alternatives, which can provide the same benefits. It follows that their reflections can be qualified as a moderate stance since they propose an approach based on a modest expansion of criminal law. They contend that "AI punishment should be avoided – not because it is incompatible with criminal law, but simply because it is a bad idea".[312]

The alternatives that they propose are several, but only two will be mentioned here. First, to create a new constructive liability crime[313] called "Causing Harm Through Criminal Uses of AI" which would fill the gap left by situations where a human agent has deployed an AI system to commit a crime (the so-called base crime), but then the agent unforeseeably commits a further (criminal) result. Notice how this would work as an alternative to the natural probable consequence model theorized by Hallevy. It implies that the agent would be liable for the more serious crime without requiring any type *mens rea*, assuming that the base crime (for which *mens rea* is instead required) "carries at least the risk of the same general type of harm as the constructive liability element at issue".[314]

They rule out the possibility of introducing new *negligent* crimes upon developers for having developed a system that foreseeably could produce a risk of harm and that of imposing this responsibility through strict liability. Second, they propose the establishment of the so-called Responsible Person regime: the authors theorize creating "a designated adjacent person" which could be punished and "who would not otherwise be directly criminally liable".[315] In his most recent work, Abbott also suggests the creation of new legal duties to "responsibly develop, supervise, or remain accountable for an AI, with liability for failing to discharge those duties".[316]

3.3.2 Lagioia and Sartor

Lagioia and Sartor use the RDS to validate their proposed theoretical framework regarding the commission of *intentional* crimes by AI systems, specifically focusing on cases where the control relation between humans and AI systems is of low intensity. This means that they only refer to situations where AI systems are not constrained to comply with the request of

312 Abbott, 2020, p. 16.
313 The authors describe them as "crimes that consist of a base crime which require mens rea, but where there then is a further result element as to which no mens rea is required". Abbott & Sarch, 2019, p. 372.
314 Ibid.
315 Abbott & Sarch, 2019, p. 378.
316 Abbott, 2020, p. 16.

users since they are not monitored, nor directed through instructions.[317] Their aim is to discuss whether AI systems can "realise crimes, respond to reasons, and be influenced by criminal norms".[318] They refer to "AI crimes" as cases in which AI systems satisfy both the *actus reus* and *mens rea* requirements of crimes.

The authors adopt a narrow definition of *actus reus*, similarly to Hallevy. Namely, they argue that it comprises of "a material aspect having a factual-external presentation" that "does not include the agent's capacity to engage in practical reasoning, guide its actions and actualize result".[319] It follows, according to this definition, that the objective requirement can be fulfilled by involuntary and unwilled action, and therefore that AI systems, be it those who act in the physical sphere such as robots, or stand-alone software, can always act in a way that is relevant for criminal law in an *actus reus* perspective. Yet, this characterization suffers of some limitations. Amongst all of them, it does not acknowledge that the act requirement has held the criminal law "hostage" for a long time now,[320] and this led to the development of a myriad of theories of action in the past decades.[321] As the discussion of the main theories of conduct of the selected legal systems will be conducted in the second part of this Chapter, the assessment will be limited to this one consideration for now.

Concerning *mens rea*, they divide their argument in two parts, one covering the *cognition* element of intent (i.e., the agent's awareness of "factual reality", which must include all the elements of the *actus reus* such as the "act or course of conduct, surrounding circumstances, and act's outcome or result element")[322], and one covering the *volition* element of intent.

As for the question of whether an AI can fulfill the *cognition* requirement, the authors give a positive answer. They define cognition in terms of "situation awareness", which can be deconstructed into three levels:

(1) Perception of the elements in the environment (i.e., the ability of an agent to detect and monitor the variables of an environment at a particular point in time);

---

317 Lagioia & Sartor, 2020, p. 3.
318 Ivi, p.5.
319 Lagioia & Sartor, 2020, p. 8.
320 Keiler, 2013, p. 43.
321 Keiler, 2013, p. 53.
322 Lagioia & Sartor, 2020, p. 9.

(2) Comprehension of the current situation (i.e., the ability of the agent to combine and interpret the information collected at the previous level, and to integrate it with pre-existing knowledge);
(3) Projection of future status (i.e., the ability of the agent to make guesses on future events).[323]

For example, a doctor needs to be aware of the patient's age, allergies, medical history, therapy, and should be able to detect her status and to interpret her symptoms.[324] The authors contend that the information methods adopted by AI systems are just "different ways of implementing the same cognitive functions" that are employed by humans.[325]

According to Lagioia and Sartor, then, AI systems, can fulfill the cognition requirement: they are capable of acquiring percepts through sensors (in the case of physical robots), through tracking activities and through messages (in the case of software agents); they can construct general images through data analysis and combination with previously stored patterns; they can make reasonable decision by computing probabilities of alternative courses of action.[326] In simpler words: "an AI is fully able to perceive its environment, comprehend it and make future projections about it".[327]

As for the *volition* requirement, specifically intent, Lagioia and Sartor adopt Bratman's conceptual framework for intent, i.e., the BDI model, as do Abbott and Sarch. Interestingly so, the BDI model is used not only to analyze human agency and cognition, but was also adapted in computer science to work as programming approach that can then be used to develop AI systems.[328] The model deconstructs intent, meant as a commitment to act, into three mental attitudes:

(1) belief (i.e., "the agent's current awareness of a situation plus any inferences it can make from them");[329]

---

[323] Ibid.
[324] Lagioia & Sartor, 2020, p. 10
[325] Ivi, p.12.
[326] Lagioia & Sartor, 2020, p. 10
[327] H. Ashton, "Definitions of intent suitable for algorithms", *Artif Intell Law*, 2022, p. 30.
[328] Lagioia & Sartor, 2020, p. 14.
[329] Ashton, 2022, p. 30.

(2) desire (i.e., the objectives/goals possessed by the agents, which may conflict);
(3) intention (i.e., the output given the agent's beliefs and desires or also "some conclusion of the agent's beliefs and desires"[330]).[331]

How does a BDI-agent work?

When an agent forms new beliefs, it proceeds to evaluate which plans have invocation conditions that correspond to its internal beliefs. Additionally, it may construct or adapt its existing plans in order to achieve its goals under the new conditions. The emerging set of plans corresponds to the agent's intentions, and each plan defines a possible course of action. Therefore, intentions refer both to an agent's commitment to its desires (the goal to be achieved through the selected plans) and its commitment to the plans selected to achieve these goals.[332]

Therefore, the authors claim that it is possible to attribute criminally relevant cognitive states to an AI system when this system implements a BDI model. In other terms, it can be argued that these systems have "awareness of the relevant facts and make intentional choices".[333]

Their solution is based on two premises. First, they rule out proving the intention of the artificial agent by inspecting its internal functioning at the time of the criminal action, since it could be unavailable, no longer retrievable, or detectable. Second, they exclude the option of obtaining proof of intent by asking the system for reports on its internal state, since the system could be a built-in "liar", i.e. it could be programmed with the capacity of falsifying the explanations.

Furthermore, the authors maintain that (human) offenders are presumed to possess intent in criminal law when (1) they had full awareness of the actions and (2) their consequences were highly likely and could have been anticipated. It follows that the same presumption could be applied to AI systems. Indeed, some of these systems can evaluate the likelihood of certain events happening and decide their course of behavior on those bases, assessing what would be the likeliest outcome of their action.

330 Ashton, 2022, p. 30.
331 Lagioia & Sartor, 2020, p. 14.
332 Ibid.
333 Lagioia & Sartor, 2020, p. 15.

Lagioia and Sartor are moderate in their approach for one fundamental reason: they argue that "criminal AI systems" will indeed require an *ad hoc* response, yet they base their conceptual framework on the liability of humans, not on the liability of the AI agent itself.[334] Let us explain this statement.

Lagioia and Sartor propose four solutions to fill the liability gap. First, limiting the tasks assigned to AI systems to not "sensitive areas" and/or their autonomy in these areas (i.e., "keeping humans in the loop").[335] Second, they focus on civil law remedies. Third, to expand the boundaries of the crimes applicable to the "humans in charge of the AI system".[336] Fourth, directly punishing the AI system. The first two solutions are deemed insufficient to address the issue at stake and fall outside the scope of this Chapter. Hence, this study will only focus here on the last two solutions, which are placed under the same category by the authors, namely "A Specific Criminal Liability for Creating and Deploying Criminal AI Systems".

Let us shift the attention now to solution number three, namely "punishing the behaviour of the user/controller who has intentionally or negligently allowed the AI system to develop criminal behaviour (e.g. by adopting an architecture that enabled such behaviour or by omitting the controls that, for example, allowed the system to evolve becoming dangerous)".[337] The proposed liability architecture is complex and therefore needs to be deconstructed in separate elements. The proposed offense would punish:

(A) the user/controller;
(B) who has allowed the AI system to develop criminal behavior;
    1 through an action; or
    2 through an omission;
(C) intentionally; or
(D) negligently.

[334] Ivi, p. 25.
[335] Ibid.
[336] Lagioia & Sartor, 2020, p. 27.
[337] Ibid.

With regards to (A), i.e., the recipient of the criminal norm, this would be the user or the "controller" of the AI systems. One could ask herself who falls under the label of "controller", as the term is very general, and the authors do not give a strict definition.

With regards to the act element of the *actus reus* (B), the conduct is described as the creation/deployment of AI systems capable of criminal conduct. This conduct could take place as an *action* (B) (1), i.e., adopting an architecture that enables AI criminal behavior, or (B) (2) as an *omission*. With regards to the latter, the examples proposed are various, such as omitting the controls which were necessary to avoid that the AI systems became dangerous, and the failure to include a "normative architecture"[338] in the system. AI systems can be considered normative agents if they follow the norms of the society in which they operate.[339] These normative constrains could prevent, on the one hand, the adoption of criminal means by the system to achieve permissible goals (e.g., engaging in fraud for maximizing profits) and, on the other, the direct pursuit of criminal goals (e.g., killing an adversary).[340]

With regards to the *mens rea* element, the conduct could be intentional (C), or negligent (D). With regards to negligence (D), the authors argue in favor of broadening the scope of recklessness to cover what they refer to as "opaque recklessness"[341] or *dolus eventualis*.[342] This form of *mens rea* would be realized using an AI system "in the awareness that it might engage in criminal activities … even when the user did not foresee that the system would engage in the specific activity".[343] In the case of the RDS example, the artist would be liable for purchase of Ecstasy by the AI system, since they were aware that it was capable of committing unlawful commercial transactions. Lastly, the authors argue that the criminal

---

338 Ibid.

339 The translation of norms into code is not as easy as it may seem when it comes to identifying which norms should be encoded in the agent. Think of the example of a person driving a friend to the hospital because she ruptured her appendix. The driver, to get her friend to the ER as fast as possible, incurs in a series of violations of the norms regulating how to drive: speeding, running through red lights, driving on a sidewalk – although carefully, avoiding accidents. Said behavior is explained by the fact that the driver / agent assigned higher priority to saving her friend's life rather than to obeying traffic laws. Think of the same situation, but with an AV. How can this behavior be encoded into the architecture of an AI system? See P. Langley, "Explainable, Normative, and Justified Agency", *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, p. 9776.

340 Lagioia & Sartor, 2020, p. 27.

341 The term was coined by K. K. Ferzan, "Opaque Recklessness", *ournal of Criminal Law and Criminology*, vol 91, Issue 3, 2001.

342 Lagioia & Sartor, 2020, p. 27

343 Ibid.

punishment of the human agents could work as a useful tool, adopting an approach which they define as "pragmatic", i.e., focused on the deterrence effect of criminal sanctions. AI systems could be programmed in a way that they recognize the benefits and the costs of inducing in behavior, which will lead to sanctions for its users (and controllers).

Finally, the authors address the fourth solution, i.e., directly punishing AI systems, in a merely speculative way. They argue that it would be possible, under a deterrence rationale, based on two assumptions, which are (1) that one day AI systems will be capable of possessing funds; and (2) that the same AI systems will be programmed as to act in a self-interested way (i.e., maximizing its profits and minimizing its losses). These systems could be deterred from criminal conduct through economic sanctions. Committing a crime, in fact, would lead to a disutility for them. It follows that "AI agents aiming to maximise their utility will refrain from engaging in criminal activities leading to expected losses (sanctions) exceeding expected benefits".[344] From a rehabilitation perspective, Lagioia and Sartor affirm that "punishment could be directed to improve systems' performance, for example by refining decision-making processes through learning or by introducing norms as constraints in the system's architecture".[345] Nevertheless, the authors drastically conclude that establishing that AI systems are criminally responsible would not be impossible, "though certainly unneeded, outlandish, and merely speculative under the present circumstances".[346]

Pagallo and Barfield summarize Lagioia and Sartor's position as follows "Lagioia and Sartor argue that the cognitive states of cognition and awareness, volition and intention, or reason responsiveness, can be attributed to an AI system in a way that is meaningful for current criminal lawyers, although no reference is necessary to human-like properties à la Hallevy's position".[347] In other words, these authors show that Abbott's Hard AI-Crimes can exist in reality, i.e., that AI systems can engage in activities that would constitute crimes if accomplished by humans. At the same time, they distance themselves from AI punishment expansionists, as they rule out direct punishment of AI systems.

3.3.3 Freitas, Andrade, Novais - Nora Osmani

[344] Lagioia & Sartor, 2020, p. 29.
[345] Ivi, pp. 28-29
[346] Lagioia & Sartor, 2020, p. 29.
[347] Pagallo & Barfield, 2020, p. 140.

These four authors were grouped together since they all rely heavily on analogies with corporate liability.

Freitas et al.[348] ask themselves whether there are any substantial differences between AI systems and corporations which would justify an exclusion of criminal liability of the former. They first focus on Hallevy's direct liability model[349] and claim that they are confident that if such a liability were to exist, it would not replace the programmer or the user's liability. They would co-exist as it happens with corporate criminal liability, where "punishment of the individuals behind the legal entities does not constitute a requirement to have the criminal punishment of the legal entities themselves".[350] Then, they ask whether an AI system could fulfil *actus reus* and *mens rea* requirements. With regards to the former, specifically the act, they believe that it is inadmissible, as they adhere to the traditional definition of "acting".

In other words, it is not sufficient to claim that this requirement would be fulfilled by a muscular (mechanical, in the case of robots) movement: doing so would result in disregarding AI systems that do not possess a physical presence, but are yet capable of harm, such as stand-alone software. In case of this latter type of AI systems, one cannot claim that the physical act is represented by an electronic impulse, either. This would be akin to stating that in the crime of defamation "the relevant act corresponds to the movement of one's tongue, mouth and vocal cords".[351] What is more, they argue that the act punished by the offense shall be "voluntary" and that this is an inquiry which shouldn't be confused with the one regarding *mens rea.* Indeed, the claim that "[t]here can be volition without *mens rea*, but the contrary is not true".[352] As a consequence, they believe that while it is possible to make a case for identifying volition in the acts of legal entities, "to plunge into the same conclusion as to AI entities' acts would arguably be precipitated".[353]

---

[348] P. M. Freitas, F. Andrade & P. Novais, "Criminal Liability of Autonomous Agents: from the unthinkable to the plausible" in P. Casanovas et al. (Eds.), *AI Approaches to the Complexity of Legal Systems. AICOL 2013 International Workshops, AICOL-IV@IVR, Belo Horizonte, Brazil, July 21-27, 2013 and AICOL-V@SINTELNET-JURIX, Bologna, Italy, December 11, 2013, Revised Selected Papers*, Lecture Notes in Computer Science, Vol. 8929, Springer, 2014.

[349] See Ch. 3.2.1.

[350] Freitas, Andrade & Novais, 2014, p. 8.

[351] Freitas, Andrade & Novais, 2014, p. 9.

[352] Ibid.

[353] Ibid.

With regards to fulfilling *mens rea*, these authors claim that it might represent a challenge which is difficult to overcome. Most importantly, adopting criminal law for the type of AI systems that are deployed now would be "rather useless and unjust".[354] Yet, the authors do not exclude that in the future AI systems could fulfil the minimum requirements requested to be considered blameworthy and lead to criminal punishment. That is, they say, an instance of the "flexibility" of criminal law.

Osmani argues in favor of shifting from and individual-centered liability model to an organization centered one when dealing with AI-caused misbehavior. She claims that the (negligent) "responsibility of manufacturers, distributors, and users of AI systems thus depend upon their capacity to understand the behaviour patterns of AI systems, the causal possibilities of AI systems' actions, and expected results".[355] Yet, these subjects might not possess the ability to predict this behavior due to how AI systems function. She believes, then, that concepts like "foreseeability" and "reasonable care" cannot constitute the elements on which one can base the imposition of criminal liability on manufacturers and sellers, "as they may simply not have the level of skill required to foresee the manner in which the harm will occur."[356] As a consequence, "standards of reasonable care may be vague. In addition, any attempt to impose responsibility on such a basis could lead to infinite liability for creators of AI systems that could obstruct the economy and innovation of AI".[357]

The result is a vicious circle. Hence, the author turns her attention to a different subject: big tech corporations. She argues that the "deep pockets of corporations",[358] i.e., those who "collect the fruits of AI deployment",[359] should be held accountable. This could be done by developing corporate criminal liability through a "correlate stream of public welfare doctrine for strict-liability offences".[360] Public welfare offenses (or also "regulatory offenses") are defined as "minor crimes that carry modest fines, although short incarceration terms may be authorized. They are understood as "regulatory" in nature – *malum prohibitum* rather than *malum in se* – and commonly address threats to "public health, public safety, public

354 Freitas, Andrade & Novais, 2014, p. 10.
355 N. Osmani, "The Complexity of Criminal Liability of AI Systems", *Masaryk University Journal of Law and Technology,* Issue 14, No. 1, 2020, p. 66.
356 Ivi, p. 67.
357 Osmani, 2020, p. 67.
358 Ivi, p. 69.
359 Osmani, 2020, p. 69.
360 Ivi, p. 73.

morals or public order". They lack of *mens rea* requirements for almost or all elements, even those who define wrongfulness.[361] Osmani argues that adopting this theory would serve two purposes. Firstly, since this doctrine was created to address the new societal threats brought about by the industrial revolution, it could be applied to the threats that are brought about by the digital revolution and could "pave the way for charging corporations with criminal offences for the harmful acts of AI".[362] Secondly, as the doctrine omits blameworthiness as a requirement for the offense, it would resolve one of the core issues of imputing criminal liability for AI misbehavior.

3.3.4 Simmler and Markwalder

According to Monika Simmler and Nora Markwalder, "robotics could prove an 'exceptional occasion for changes to the law and legal theory".[363] Consequently, criminal law scholars should not wait and not address technological developments as it could result into yet another late reaction. As other scholars analyzed in this Chapter, Simmler and Markwlader focus on the liability of *robots*, which they define as "machines (mostly with sensory-motor functions), which are built to enhance human possibilities for action".[364]

The focus of their inquiry is on a scenario in which "a robot 'commits' a crime, because it has developed its own momentum due to its artificial intelligence"[365] and the momentum, which is "predetermined and depends on the programming […] cannot be traced back to a single programming operation".[366] Such a scenario recalls discussions on the existence of free will, i.e., situations "in which we can assume that every action has been caused and determined by something somewhere, but in which we attribute this action to a person as 'their own' and as an 'act of (free) will', because we cannot trace and explain the exact process of causation".[367] The fact that this article is written by criminal legal scholars

361 D. K. Brown, *Public Welfare Offenses*, in Dubber & Hörnle (Eds.), 2014, p. 864.
362 Osmani, 2020, p. 73.
363 M. Simmler & N. Markwalder, "Guilty Robots? – Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence", *Criminal Law Forum*, No. 30, 2019, p. 4.
364 Ivi, p. 5.
365 Ivi, p. 8.
366 Ibid.
367 Simmler & Markwalder, 2019, p. 9.

and focuses extensively on discussing the potential criminal liability of the robot, and not on humans, makes this article noteworthy.

Taking a closer look at the arguments made in the paper, its primary concern is the "fatal question of criminal law", i.e., culpability.[368] The authors contend that there are two main problems with attributing of criminal liability to robots: first, "(the assumption of) human free will as the foundation of the traditional theory of criminal responsibility";[369] second "the question whether intelligent robots could be recognized as a subject of criminal law and therefore concerns the very basic concept of legal personhood".[370]

With regards to the first issue, the authors contend that robots could never fulfil the traditional definition of the fault principle, regardless of the advancement in their technology, since a robot is not a "person with free will" (meant as the capacity of human beings of being autonomous persons who have the ability to choose between alternative actions, i.e., to consciously decide not to act in accordance with the expectations placed upon them).[371] Taking into account the relevance of the impact of neuroscience on the issue of free will, the authors argue that "it does not make sense to succumb to the vortex that is the debate around the ability to choose one's action and to a scientific analysis of determinant variables",[372] since human free will is a byproduct of the social construction of reality. It follows that the biophysical nature of free will loses relevance for the purposes of criminal law, also when it comes to robots' liability.

If one discards the traditional notion of human free will, then, the authors contend, it must look at it as an attribution "in the socially system",[373] rather than as a set of "objectively and individually confirmable characteristics".[374] And if it is society that attributes freedom – and therefore attributes criminal liability – as a social fact, based on its perception, then it is only "a question of time until humans experience this autonomy not just as determined and programmed and until they attribute robots the respective 'capacities".[375]

---

368 E. Hafter, *Lehrbuch des Schweizerischen Strafrechts, Allgemeiner Teil*, 2nd Ed, Springer, 1946, p. 101.
369 Simmler & Markwalder, 2019, p. 10.
370 Ibid.
371 Simmler & Markwalder, 2019, p. 11.
372 Ivi, p. 12.
373 Ivi, p. 14.
374 Ibid.
375 Simmler & Markwalder, 2019, p. 15.

With regards to the second issue, the authors refer to Gless and Weigend and claim that they subscribe to a traditional understanding of the question of personhood in criminal law, which "neglects the fact that the concept of the subject or of personhood in criminal law is constructed in social reality and does not necessarily refer to biophysical categories". These authors, it is argued, "adopt a German perspective".[376] Indeed, as it will be explained in a later moment, Gless and Weigend contend that robots cannot qualify as addressees of criminal law since they cannot conceive themselves as morally responsive agents and because they cannot understand the concept of retributive punishment.[377]

Simmler and Markwalder claim that a more modern conception of person in criminal law revolves around the agent's capacity to "*disappoint normative expectations and on the possibility of attributing actions to this agent in the social system"*.[378] As such, criminal capacity – which will be dealt with in detail at Chapter 5 – is an "artificial concept developed by and for the purpose of observers of the social system, which summarizes existing expectations".[379] As an artificial concept which is related to "a specific society in a specific time",[380] it is subject to changes, i.e., it could be also applied to robots.

Corporate criminal liability is the proof of the fact that non-human entities can be subjects of criminal law. indeed, society possesses certain expectations towards corporations since they are agents in the social system (i.e., they have an influence on everyday life and on social interaction), as such, corporations are deemed capable to destabilize norms and disappoint expectations.[381] According to the authors, robots are not there yet: even though it might be possible to deem robots "guilty" for a criminal offense, they cannot be treated as perpetrators, since criminal responsibility must be ultimately read through the glasses of criminal punishment.

The authors rely heavily on Jakob's theories of criminal law, and claim that "[t]o deem someone 'guilty' means nothing else than that we impute a fault, the disappointment of a normative expectation, to a person".[382] It's society's way to "externalize the conflict, resolve

---

376 Ivi, p. 18.
377 Gless & Weigend, 2014, p. 570; Gless, Silverman & Weigend, 2016, p. 412.
378 Simmler & Markwalder, 2019, p. 17.
379 Ibid.
380 Ibid.
381 Simmler & Markwalder, 2019, p. 19.
382 Ivi, p. 25.

it, and stabilize the norm put into question and thus secure the survival of said norm".[383] As a consequence, whether robots are criminally liable or not depends exclusively on whether society wants to attribute them the role of "rule-breakers" or not and on whether criminal punishment would resolve the conflict which results from the rule breaking. Criminal punishment fulfils its aims only if it is directed as subjects which have received such an attribution from society. And, since it is a *sociological fact* that robots have not (yet) acquired such personhood, their criminal punishment must be excluded.

Finally, the authors shortly reflect on how robots could practically be punished in the future: they assume that a fine could be imposed, if future robots will be able to earn and lose money; and that robots could be reprogrammed or be inflicted a non-specified form "evil", which would have consequences for the self-learning system.[384]

Their final claim is that

> criminal responsibility of robots is possible if it is in accordance with the system, ie with the function of criminal law, if it is useful and necessary in the context of the stabilization of norms. This is the case only if robots have the necessary requirements of personhood and if they are thus attributed the capacities that are inherent in that concept. […] A possible attribution of criminal responsibility to robots is thus a process subject to the function of the system. If these requirements are met, robots can therefore be subject to criminal responsibility.[385]

They then identify three main alternative rules:

> (1) to insist on the traditional approach to criminal responsibility, which rests on freedom of choice, on the ideal of the autonomous human and on the exclusive application of the concept of personhood to humans and to defend this idealistic concept against all attack and even in the face of increasing instability. This route would lead to abandoning the possibility of attributing criminal liability to robots

[383] Ibid.
[384] Simmler & Markwalder, 2019, p. 28.
[385] Ivi, p. 29.

> altogether and would lead to concentrating on the programmer or operator of the robot for questions of criminal responsibility;
> (2) to adopt a functionalistic approach to criminal responsibility, in the sense that the concept of criminal responsibility expresses a reproach for a socially visible lack of compliance, which has the potential to destabilize norms. Such an approach is not based on recognizable human free will or freedom of choice, rather on the purposes of criminal law. It follows that a robot would be deemed liable if it would actually be experienced as an "equal" in the sense it would be constituted as addressees of normative expectations in social interaction like humans or corporate entities are today. They qualify this route as the most realistic route for the future;
> (3) to take advantage of opportunity of the discussions on the concept of criminal law that are caused by technological advances to overcome and to rethink the traditional approach to the concept of criminal responsibility. They qualify this option as the most difficult, yet most promising, route to pursue.[386]

To conclude, Simmler and Markwalder represent a very interesting case study for this research. First of all, differently from most of the moderates, they are criminal legal scholars and not legal philosophists. Second, and most importantly, we could qualify them as "border line" moderates, since they make a very strong case for holding robots liable, only to then exclude it for the time being.

3.3.5 Mihail Diamantis: the Corporate Mind and Body Approach

Diamantis, too, suggests a theory for AI liability deeply grounded in corporate liability. Yet, he adopts an innovative perspective compared to Osmani and Freitas et al.: he claims that corporations could act not only through their employees but also through their algorithms. In a nutshell, this entails that algorithmic action should be considered as corporate action and that therefore algorithms could become an "extension of the corporate person".[387] This, he believes, is the "path of least resistance" for addressing algorithmic injury.

---

386 Ivi, pp. 29-30.
387 M. Diamantis, "Algorithms acting badly: A Solution from Corporate Law", *GEO. Wash. L. Rev.*, Vol. 89, 2021, p. 809.

The discussion conducted by Diamantis is extensive and requires familiarity with concepts of corporate criminal liability which stems from U.S. law but aspires to have an application beyond these borders. For these reasons, Diamantis' theory will be analyzed in Para. 6.4, which is dedicated to the intersections between AI liability and corporate criminal liability.

3.4 SKEPTICS

The skeptics front of the academic debate on AI and criminal liability is composed by those who dismiss the idea of punishing AI from the start as a conceptual confusion: AI systems cannot possess a guilt mind, and they cannot perform a guilty act. In other words, they reject the idea of directly punishing AI systems based on classical criminal law principles.[388] Authors who advocated for minor changes in the traditional categories of criminal law were also placed in this category.

3.4.1 The Italian Approach

The issue of criminal liability for AI misbehavior has been gaining more and more momentum in the past few years amongst Italian criminal legal scholars. At times, it seemed as if everyone was jumping on the AI-bandwagon. Nonetheless, this did not entail that the contributions to the debate were not of value. Most of the articles that were published are in Italian, while quoting authors who wrote in German and in English. This entails that while Italian-speaking scholars took advantage from the ongoing international debate, they did not actively contribute to it. This pinpoints one the strengths of this research: crossbreeding of legal doctrine.

Different authors were included in the same section under the label "The Italian Approach" for various reasons. To begin with, almost the totality of the analyzed authors denies categorically the possibility of imposing criminal liability on AI systems.[389] This might

[388] Abbott, 2020, p. 112.; Abbott & Sarch, 2019, p. 327.

[389] E. Lo Monte, "Intelligenza artificiale e diritto penale: le categorie dommatiche alla prova del futuribile", in F. Basile, M. Caterini & S. Romano (Eds.), *Il sistema penale ai confini delle hard sciences*, Pacini Giuridica, 2020; F. Basile, "Intelligenza artificiale e diritto penale: qualche aggiornamento e qualche nuova riflessione", in Basile, Caterini & Romano (Eds), 2020; G. Rizzo Minelli, "Quando l'autore del reato è un

be explained by the fact that the Italian criminal system is deeply rooted in the concept of culpability as personal punishment, following the brocard *nullum crimen sine culpa*. Further, many authors reason in terms of "risk society" and of *"diritto penale del rischio*".[390] Moreover, differently from other sides of the debate represented in this Chapter, the Italian front is characterized by overarching analyses. In other words, Italian scholars seem to prefer cherry-picking the most trivial issues regarding the impact of AI on criminal law, rather than to develop general theories. This entails that the best approach for the literature review of this side of the debate is to combine the different voices.

One of the most relevant analyses in the Italian arena is the one conducted by Alberto Cappellini. Cappellini starts his reflection by reformulating an ancient brocard:[391] *machina delinquere (et puniri) non potest*.[392] Indeed, Italy adopted an administrative (quasi-criminal) liability for corporations in pursuant to legislative decree 231 of 2001. According to leg. decree 231/2001, if a criminal court ascertains that a 'qualified' subject (i.e., a CEO or a high-ranking manager) committed one of the offenses listed in leg. decree 231/2001, the same criminal court can then sanction the corporation in whose interest, or for whose benefit, the crime or crimes were committed. The Italian model will be further explained in Ch. 4.b.4.1.

Capellini then asks whether this axiom still applies, considering the most recent evolutions in AI technology. As we know, some, such as Hallevy,[393] reply affirmatively. Cappellini qualifies Hallevy's reconstructions as "unacceptable to date to the point of

---

robot: tra vecchi modelli imputativi e nuovi possibili paradigmi di responsabilità" in Basile, Caterini & Romano (Eds), 2020; S. Riondato, "Robot: talune implicazioni di diritto penale" in P. Moro & C. Sarra (Eds.), *Tecnodiritto. Temi e informatica e robotica giuridica*, Franco Angeli, 2017; V.C. Talamo, "Sistemi di intelligenza artificiale: quali scenari in sede di accertamento della responsabilità penale?", *Il Penalista*, 2020; R. Borsari, "Intelligenza Artificiale e responsabilità penale: prime considerazioni", *MediaLaws*, Vol. 3, 2019. Cecilia Cavaceppi argues that state-of-the-art-AI should fall under the current framework for (administrative) corporate liability provided in Italy by decree law 231/2001 whenever it is the product of a programming firm offering its services to third parties. C. Cavaceppi, "L'intelligenza artificiale applicata al diritto penale", in G.Taddei Elmi & A. Contaldo, *Intelligenza artificiale-Algoritmi giuridici: Ius condendum o fantadiritto?,* Pacini Giuridica, 2020. In this scenario, Lagioia, Sartor and Pagallo represent a *unicum*, probably because they have a background in legal philosophy, rather than in criminal law.

390 For example Piergallini, Panattoni, and Salvadori.

391 The original brocard relates to the liability of corporations and reads *societas delinquere non potest*. Admittedly, Franz von Liszt coined the phrase in 1881.

392 A. Cappellini, "Machina delinquere potest? Brevi appunti su intelligenza artificiale e responsabilità penale", *Criminalia*, 2018.

393 For an analysis of Gabriel Hallevy's work, see section 3.2.1. of this Chapter.

appearing provocative, to say the least, in the eyes of many criminal scholars of continental education".[394] The author then puts forward a number of critics to Hallevy's theory. The following paragraphs will mention two of them.

The first criticism is that Hallevy did not prove that AI systems are capable of culpable conduct. Cappellini holds that what Hallevy proved is just an appearance of intent, a mere "sleight of hand". According to Cappellini, then, it might be true that AI systems act with a certain level of discretion, which allows them to adapt their behavior to the external reality in the pursuit of their aim. Yet, this is not enough for criminal law: the choices taken by the AI systems are not the result of a capacity of self-determination. These systems are not capable of willfully choosing to perform an illegal act. For these reasons, they cannot be reproached; hence, they cannot be criminally liable.

The second criticism is that the analogy with corporate liability is fallacious. Cappellini argues that corporations, unlike AI systems, always lack a physical *persona* in the real world. They only exist in the juridical and social world, where instead (some) AI systems, such as robots, not only exist in the physical world, but they are also capable of detaching from their creators. The same cannot be said about corporations, which always play as puppets in the hands of human puppeteers. For these reasons, says Cappellini, imposing criminal liability on corporations has a direct effect on the "wallets" of said human puppeteers, where instead directly punishing the AI system would have no effect on the human being behind the curtains who has lost control of the machine. Even if it did, it would prove useless, because she would not be able to influence the behavior of the (uncontrollable) AI system.

Fabio Basile poses several open questions, starting from Cappellini's *machina delinquere non potest* brocard.[395] If criminal responsibility is personal, meaning as relating to a *persona*, can we then really assimilate men to machines? Is there not something "more" about being human? The questions remain unanswered, yet the author's concern is palpable. Others claim that the logical step required to negate the competence of criminal law to regulate AI systems lies in the fact that even the most advanced systems are nothing but machines which

[394] Cappellini, 2018, p. 11.

[395] Basile in Basile, Caterini & Romano (Eds.), 2020, p. 34.

elaborate data through mathematical operations, based on given instructions.[396] Reasoning on the purposes of punishing AI systems is, henceforth, regarded as preposterous.[397]

According to Manes, the most suggestive scenarios which deserve discussion are two. First, situations where the criminal conduct is the autonomous result of a software, which can be assisted by the inert and only eventual human presence. Second, situations where the criminal conduct is the result of shared action between humans and AI systems. Thus, he asks himself on whom responsibility for the causation of an event shall be ascribed in these situations. On the person who has generated the source of risk by designing the software (especially in cases of self-learning algorithms)? On the person who has actualized that risk by manufacturing and putting the system in the market? Or on the person who has concretely managed that risk by using it, or perhaps by cooperating with it (in cases where it would be possible to ascertain a sort of *culpa in interagendo*)?[398]

Magro investigates whether we can consider AI systems as possessing artificial free will or if their autonomous behavior is just a deviation from the original project caused by accidental factors.[399] It does not really matter, she claims, to define which level of freedom AI systems can display. Criminal liability is not bound to the adoption of a univocal definition of freedom for all agents. What matters is to understand whether we can consider AI systems *responsible* and this relies on which conception of responsibility we adopt.

Consistently with the rest of the Italian front, Magro makes a strong case against holding AI systems directly liable and she also rejects the analogy with corporate criminal liability. Casting aside the option of considering AI systems as (criminal) agents, and focusing on the human agents (the "frontman"),[400] in her earlier work she proposes the creation of a new legal concept, "*colpa da programmazione*"[401] or "fault or negligence by programming".[402]

---

396 Lo Monte, 2020, pp. 66-67; Riondato, 2017, p. 92.

397 Basile in Basile, Caterini & Romano (Eds.), 2020, p. 76.

398 V. Manes, "L'oracolo algoritmico e la giustizia penale: al bivio tra tecnologia e tecnocrazia", *Discrimen*, 2020, p. 3.

399 M. B. Magro, "Biorobotica, robotica e diritto penale", in D. Provolo, S. Riondato & F. Yenisey (Eds.), *Genetics, Robotics, Law, Punishment*, 2014, p. 516.

400 Ivi, p. 8.

401 M. B. Magro, "Decisione umana e decisione robotica. Un'ipotesi di responsabilità da procreazione robotica", *La legislazione penale*, 2020, p. 8.

402 M. B. Magro, "Biorobotics, robotics and criminal law: some hints and reflections", *Percorsi costituzionali*, Fasc. 1-2, 2016, p.9.

As it will be demonstrated, this concept is recurring in Italian legal doctrine and is often discussed in relation to duty of care obligations.[403] It has not been cleared, yet, whether this type of negligence would be an example of what Italian criminal legal doctrine refers to as '*colpa generica*' (referred to also as "unconscious negligence"),[404] i.e., negligence based on the violation of general (unwritten) standards of care; or – more likely – whether it would be built based on written (technical) standards (*colpa specifica*).

Then, she argues, we should create a new circumstance which excludes liability whenever programmers put certain measures in place (which would mean, in practical terms, to make sure that robots abide to Asimov's laws).

Generally speaking, imputing liability for AI systems misbehavior in these situations would require an operation which is opposite to the one of imputing liability for corporate misbehavior. When it comes to corporations, the action of a human agent makes a non-human agent liable. When it comes to AI systems, the action of a non-human agent acts makes the human agent liable.[405]

In her later work, she also claims that programmers or designers of the software would not be able to escape negligence liability by claiming that the harmful conduct of the AI system was not predictable due to its autonomy.[406] Actually, according to the latest evolution of negligence in Italian jurisprudence, negligence requires only the abstract foreseeability of a general risk, rather than the foreseeability of concrete and specific harmful event. It implies holding a subject liable for not avoiding worst case scenarios or catastrophes, i.e., events that cannot be predicted or avoided by the subject, such as earthquakes. This type of negligence is also referred to as "*colpa eventuale*".[407] It is self-evident, then, that it would be impossible for the programmer to deny that she was not aware of having created independent and self-learning machine.

Panattoni contends that the causation of harmful events by AI is another instance of our risk-based society, which mandates an anticipation of the protection provided by criminal

---

[403] Manes, 2020, p. 4.
[404] Keiler & Roef (Eds.), 2019, p.198.
[405] Magro, 2014, p. 514.
[406] Ivi, p. 516.
[407] G. Civello, *La "colpa eventuale" nella società del rischio. Epistemologia dell'incertezza e "verità soggettiva" della colpa*, Giappichelli, 2013.

law according to the level of danger of the action. This is reflected also in the EU's approach to regulating AI, which is, indeed, risk-based. [408]

Salvadori[409] argues that any attempt to impose criminal liability on AI systems is destined to wreck against the cliff of article 27 of the Italian Constitution,[410] which establishes that criminal liability is personal. This principle entails that the offense must be reconducted to the offender through a phycological nexus and that the offender must be blameworthy. It would not be feasible to do the same with AI systems. The issue, then, becomes one of distributing liability amongst human and artificial agents.

Interestingly so, he claims that the scenario involving a human operator who decides to create a software for malicious purposes does not fall into the perpetrator-by-another model. Indeed, differently from cases of indirect perpetration, in the former case the agent would be liable for an act that she put in place, as she used a tool to commit a crime, where instead in cases of indirect perpetration the main perpetrator is liable for the conduct of another perpetrator.[411] With regards to the realm of *colpa*, Salvadori, as Magro before him, asserts that a defect in the functioning of the software could be pinned upon the developer and/or the programmer based on a *colpa di programmazione,* which means not having predicted the possibility of grave incidents. Indeed, these subjects have a duty to monitor how the software works and to update it whenever necessary.

In conclusion, he believes that solution which will be adopted with regards to criminally regulating AI systems is that of creating a sphere of admissible risk (*Erlaubnis Risiko*), i.e., a level of risk that is tolerated by society and that is based upon specific codified duties imposed on the humans involved in the AI-production chain. Moreover, he argues in favor of the establishment of a set of administrative sanctions in case the AI product does not satisfy certain technical precautional measures established *ex ante*. Finally, he suggests the creation of legal duties to encode principles in the algorithm starting from the programming phase.[412]

---

[408] B. Panattoni, "Intelligenza artificiale: le sfide per il diritto penale nel passaggio dall'automazione tecnologica all'autonomia artificiale", *Dir. Inf.*, Vol. 2, 2021, pp. 331-33..

[409] I. Salvadori, "Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale", *Rivista Italiana di Diritto e Procedura Penale*, No. 1, 2021.

[410] Ivi, p. 98.

[411] Salvadori, 2021, p. 101.

[412] Ivi, p. 116.

Piergallini notices how criminal law is at discomfort when confronted with AI systems, almost as if it were an old tool.[413] Due to AI, the way we ascribe liability is undergoing a crisis and we are forced to ask ourselves the level of risk (*"margine di sicurezza*") that our society is willing to tolerate. This, he argues, is strictly a political choice that precedes the one of ascribing liability, since it regards how to regulate said risks preemptively. Should the "creator" of the machine be blamed because she "predicted the unpredictability" of her creation?

His article is divided into two. The first part deals with situations in which there is a human who is, somehow, "in control". Here the author provides a thorough analysis of these issues from the perspective of product (criminal) liability which focuses on two case studies: AVs and robotic surgery. The following paragraphs will focus on the first.

Piergallini attentions level 1, 2 and 3 of driving automation[414] and argues that, with regards to the subjective element, the liability constructs that we know of seem to hold. In other words, the liability for an accident caused by these types of cars will always fall on the driver, who has a special duty of care. What happens, though, when the accident is caused by a defect in the production of the car? This case, he argues, triggers the applicability of product liability. The defect could be the result of a faulty design, a faulty development or a faulty construction of the car. Criminal law, then, faces the impossible mission of identifying (beyond a reasonable doubt) the liable human agent in a huge production chain conducted by complex organizations, *mega-apparati*, which often also work together. In conclusion,

---

413 C. Piergallini, "Intelligenza artificiale: da 'mezzo' ad 'autore' del reato?", *Rivista italiana di diritto e procedura penale*, Vol. 4, 2020, p. 1746.

414 The levels have been defined by the SAE (society of Automative Engineers) International in the J3016 standards. They are: Level 0: No Driving Automation; Level 1: Driver Assistance; Level 2: Partial Driving Automation; Level 3: Conditional Driving Automation; Level 4: High Driving Automation; Level 5: Full Driving Automation. With the first 3 levels of automation the driver is driving, even if her feet are off the pedals and she is not steering. The driver is also constantly supervising the driver support features and must steer, break or accelerate to maintain safety. With the level 3, 4 and 5 automation the driver is not driving, even if she is seated in the driver's seat. Level 3 implies that when the feature requests it, the driver must drive. That is, she can only control the car in situations of emergencies. Level 4 and level 5 do not require that the driver takes over driving. The difference between level 4 and level 5 is that with level 4 automation (equally to level 3) the automatic features can only drive the vehicle under limited conditions and will not operate if these conditions are not met (i.e., if there are extraordinary circumstances the car will stop in order for the driver to take control) where instead with level 5 automation the vehicle can be driven under all conditions. SAE, 2021.

Piergallini argues that in these cases the central role should be the one of tort law, not criminal law, to avoid incurring into hypothesis of strict liability.

The second part of the article deals with cases where there is no human in control of narrow AI systems, as these act in a fully unpredictable manner. With regards to causality, he discusses whether AI behavior could constitute a factor which breaks the causal chain and, as a consequence, excludes the liability of the programmer. He answers negatively, since the maker of the algorithm knowns that she is creating a new risk, as she is creating a machine in the knowledge that she won't be able to it. The true breaking point, he believes, is *mens rea*, specifically negligence. Returning to the question asked above, Piergallini argues that one cannot blame a programmer on the grounds of negligence because she *predicted* the *unpredictability* of the system. Predictability of risk, he claims, is the DNA of negligence. Piergallini, then, focuses on a critique of Hallevy's argument. Piergallini negates that an AI can be directly criminally liable. He labels Hallevy's strictly material conception of conduct as "rough".[415] Likewise, he strongly refutes that one can consider AI systems as capable of culpable conduct: there is no space for strict liability in our criminal legal system, he asserts. The Italian front stands united.

What should be criminal law's role, then, if any? According to the author, criminal law could work as a "cooperative compliance" tool, in a perspective of co-regulation between hard and soft law. On the one hand, it could impose on corporations a duty to share their know-how on how to evaluate and prevent AI risks. On the other, *post damnum*, authorities could impose the obligation – it is not specified on whom exactly – to reprogram or to deactivate the system. Non-compliance with these orders could lead to criminal sanctions. In conclusion, Piergallini believes that we should be wary of elevating criminal law as the primary answer to the numerous questions that are raised by AI.

3.4.2 Ugo Pagallo

Regardless of his background as an Italian scholar, Ugo Pagallo's view is analyzes separately from the other part of the Italian front. The reasons are threefold: first, Pagallo has a background in philosophy of law and not in criminal law. Second, his work is more extensive

[415] "*ruvida"*, Piergallini, 2020, p. 1765.

than the one of the single authors in the "Italian Approach" sub-category.[416] Third, he writes mainly in English.

Pagallo, in his 2013, monograph takes a strong stance against attributing criminal liability to AI systems.[417]

Concerning the subjective element of a crime, he claims that there is no such thing as a "robotic *mens rea*"[418] since robots lack of the necessary pre-requisites: self-consciousness, free will and moral autonomy. He acknowledges that there are some authors[419] who argue in favor or a strong ontological stance, that is, they believe that the advancement of AI technology will produce machines that will decide autonomously in a way that is similar to human-like decision making. These same authors reject the objection that AI systems cannot be moral agents since they are just programmed machines. We, as humans, should not take comfort in the axiom that we are not programmed, while artificial agents unequivocally are.[420] Pagallo refutes these arguments: if we were to accept that a kind of human-like generation of AI system would come to life, we would also have to conceive a *mens rea* which would be "rooted in the artificial mind of a machine capable of a measure of empathy, or a type of autonomy, affording intentional actions".[421] From this also follows that "lawyers should be ready to take seriously a whole set of new offences such as robot revolutions, rebellions, robberies and so forth".[422] Matters for science fiction writers rather than for legal experts, Pagallo claims.

Concerning the objective element of a crime, in his earlier work of 2013 Pagallo labelled efforts to assert that AI systems could lead to a new set of *actus rei* as "Hollywood-

---

[416] He authored two books together with Woodrow Barfield: *Research handbook on the law of artificial intelligence*, Elsevier, 2019; and the *Advanced Introduction to Law and Artificial Intelligence*, 2020. He is also the author of the book *The Laws of Robots. Crimes, Contracts, and Torts* (2013) and of a vast number of publications and Chapters in edited books. Specifically, on the topic of AI accountability and criminal liability, he authored: "The Adventures of Picciotto Roboto", 2011; "Killers, Fridges, and Slaves", 2011; "AI and bad robots", 2017; and "From automation to autonomous systems: A legal phenomenology with problems of accountability*",* 2017.

[417] Even though his reflections are based on robots, i.e., AI systems with a physical presence, we believe that they can be extended to the generality of AI systems.

[418] Pagallo, 2013, p. 50.

[419] Chopra & White, 2011, p.77.

[420] Ivi, p. 176.

[421] Pagallo, 2013, p.76.

[422] Ibid., p. 76.

style approaches".[423] His attitude changed in his later work, "The Research Handbook on Law and Artificial Intelligence", namely in the Chapter "The impact of AI on criminal law, and its twofold procedures" written with Serena Quattrocolo. Here the authors introduce the example of Vital, a UK made robot which was appointed as board member of a venture capital firm in Japan (Deep Knowledge) to predict successful investments.[424] Let us imagine now that a wrong evaluation of Vital lead to a lack of capital increase and therefore to the fraudulent bankruptcy of the corporation. The authors contend that "humans could be held responsible only for the crime of bankruptcy triggered by the robot's evaluation, since the mental element requirement of fraud would be missing in the case of the human members of the board".[425] Therefore, the crime of fraudulent bankruptcy could be charged only upon the corporation and eventually upon the robot. This option would require, though "that most legal systems should amend themselves, in order to prosecute either the AI system as the criminal agent of the corporation, or the corporation as such".[426] Another case presented as an example is the result of reversing the usual perspective on the "perpetration-by-another" liability model: what if humans were the innocent agents or tools of an AI's bad decision? As the authors rightly observe, this scenario could lead to a new type of *actus reus* which does not require any level of *mens rea*.

Moving to the criminal liability of the human, Pagallo focuses on how AI systems might affect the *mens rea* of an individual. This research will only analyze the discussion on intent and negligence, as the other issues described by Pagallo (namely those of humans who commit a crime against the AI system) fall outside the scope of this research. In other words,

> matters of design (actus reus) and human culpability (mens rea) concerning the criminal field of the laws of robots, are more urgent than the current debate on new forms of (weak or even strong responsibility for) crimes committed by humans against

[423] The authors mention two examples: the "Robot Kleptomaniac" and Robbie CX30. Pagallo, 2013, p. 76.

[424] Pagallo U. & Quattrocolo S., "The impact of AI on criminal law, and its twofold procedures", in Barfield & Pagallo (Eds.), 2019, p. 385.

[425] Ivi, p. 404.

[426] Pagallo &Quattrocolo, 2019, p. 404.

their autonomous machines, in addition to the moral agenthood and legal personhood of robots.[427]

Pagallo first analyzes cases where the robot is designed to commit offenses, i.e., "Criminal Robots by Design". The authors identifies two scenarios:

(1) the robot is used to carry out existing kind of crimes through new robotic devices;
(2) the robot is used to carry out novel offences.

Focusing on option (2), Pagallo then identifies two sub-scenarios:

(2)(a) An individual sends or activates the robot to commit a crime (crimes of intent);
(2)(b) The reasonable individual fails to guard against foreseeable harms (crimes of negligence).

Scenario (2)(a) can be solved by adopting the perpetration-by-another liability model to identify the liable human being amongst three candidates: the programmer ("the evil designer"), the manufacturer ("faulty producer") or the ("criminal") user.[428] This model, instead, proves useless for scenario (2)(b), i.e., "cases where criminal liability hinges on negligence or lack of due care, rather than the blameworthy mens rea of designers, producers or users of robots".[429]

In these circumstances, Hallevy's natural-probable-consequence model comes into play. As it was already stated,[430] this model covers both situations where the AI system was programmed to commit an offense and then deviates from the plan, and situations where humans had no intent to commit a crime but were negligent when designing, constructing, or using an AI system. Concerning the former, Pagallo correctly states, "[i]n most legal systems, programmers, manufacturers or users of such robots would be liable for the additional crime, regardless of the unpredictability of the machine's behaviour, as it occurs

[427] Pagallo, 2013, pp. 54-55.
[428] Ivi, p. 70.
[429] Pagallo, 2013, p. 71.
[430] See Sec. 3.2.1.

with the liability model in accomplice responsibility case".[431] Concerning the latter, Pagallo provides an interesting suggestion: "since robots are machines capable of learning and adapting to changes in the environment, they are unpredictable. So, they will give rise to a new set of legal issues centered around how humans treated the machine, rather than the ways in which the machine was designed and constructed".[432] Based on this, one could formulate the following observation: since AI systems also learn from the interactions they have with their surroundings, and since their surroundings also include input derived from humans, it might become relevant to evaluate how humans treated the machine and how that related to the commission of crime by it.

Pagallo ends his discussion by introducing the concept of failures of causation. His reflections on causality are distinctive in the academic legal discourse on the topic.

Without a doubt, he claims, "crucial criteria for selecting from the entire chain of events the specific condition, or the set of conditions, that best explains a given outcome, would be challenged by the unpredictable behaviour of these machines and the complexity of network-centric applications".[433] He identifies two main issues, taking as an example an hypothetical intelligent program handling air traffic control, which has the purpose of avoiding issues such as ground damage, air-to-air collisions, communication interferences, piracy, environmental concerns; and which works in interaction with both manned and unmanned aerial vehicles. Firstly, he claims that "it seems problematic to aim at determining the types of harm that may supervene with the functioning of such a complex processing system".[434] Secondly, "the traditional idea of the reasonable person may fade away, since the duty of individuals to guard against foreseeable harms is challenged by the growing autonomy of robotic behavior and cases where no human would be accountable for the unforeseen results of the "machine intelligence's pathology".[435] Pagallo's proposed solution for failures of causation is a type of legal responsibility, which "is not established *ex ante*, such as strict liability, nor excluded *a priori*, such as immunity clauses".[436] This model entails that the criminal liability of humans should be established by courts "on the basis of the probabilities

---

431 Pagallo, 2013, p.72.
432 Ibid.
433 Pagallo, 2013, p. 74.
434 Ibid.
435 Pagallo, 2013, p. 75.
436 Ibid.

concerning how robots work through their on-board decision-making controllers, automatic recovery functions, communication devices, etc.", that is, the focus should be on the "scientific meaning of the machine's behaviour".[437] In order to avoid the overload of information which could result from this operation, the author argues that it would be useful to look at how the same concepts of causation and reasonable foreseeability are regulated in the field of contracts.

To conclude, in their most recent handbooks, Pagallo and Barfield directly refer to Lagioia and Sartor and claim that they too accept as true that "AI can engage in moral and legal reasoning"[438], i.e., AI systems can be built with a normative architecture and can engage in knowledge representation and reasoning (that is, "the ability to represent norms and/or values, and reason with them")[439]. This assumption can lead to some form of "criminal responsibility"' which is, nevertheless, different from full legal personhood of AI systems on the one hand, and from Hallevy's "human-like" assumptions on AI-awareness.[440]

Lagioia and Sartor's theory is based on the claim that AI could respond to the threat of legal sentencing. However, Pagallo and Barfield identify two flaws in Lagioia and Sartor's account of criminal accountability of AI. First, a technological issue: "we are likely far away from AI technology with such requisites as consciousness and moral understanding",[441] which are required for the AI systems to be the subject of criminal sanctions and to be deterred and rehabilitated by it. Second, a moral issue. Whether it is true that an AI system can be considered a source of good or evil, and, as such, it could be considered as an accountable agent, it is not clear to the authors how the status of AI as an accountable agent in criminal law would complement the same status in moral theory or in contracts and corporate law. How would the expansion of accountability of AI to criminal law improve current regulations? The question stands unanswered.

Finally, Pagallo and Barfield agree with the fact that AI could have some form of criminal responsibility in the sense that they could be targets of criminal law norms due to their capacity of possessing, on the one hand, a normative architecture and, on the other hand, reason-responsiveness. Yet, they draw a net distinction with Hallevy's standpoint: AI

---

437 Pagallo, 2013, p.77.
438 Pagallo & Barfield, 2020, p. 140.
439 Lagioia & Sartor, 2020, p. 1.
440 Barfield & Pagallo, 2020, p. 140.
441 Ivi, p. 141.

systems do not possess human-like awareness and they should not be considered as right bearers. Similarly to Diamantis' position, their suggestion is to deal with issues of "liability and distributed responsibility in complex AI ecosystems" through the teachings provided by forms of accountability for AI in business and corporate law: "[t]he criminal liability of AI, to be effective, lies first in its assets, after all".[442] In other words, "we should follow the money first".[443]

3.4.3 Dafni Lima

Lima's paper,[444] whether not as extensive as other workings examined in this Chapter, deserves attention. Indeed, Lima puts forward a set of questions contesting that an AI system could fulfill the *actus reus* requirement of a criminal offense.[445] While quoting American[446] and German[447] constructs of criminal liability, Lima notices that "it seems that concepts like bodily movement (or failure thereof) that are voluntary, extroversive, and socially meaningful in a way that is relevant to criminal law are essential aspects of acting".[448]

According to Lima, in a similar stance as the one taken by Gless, Silverman, and Weigend,[449] AI acts could hardly be categorized as instances of *actus reus*, since they do not have social relevance and are not voluntary. With regards to the latter, she contends

Voluntariness in this sense implies the ability to act otherwise, and an agent that is programmed to choose A when it encounters B is not necessarily choosing. Thus AI agents

---

442 Barfield & Pagallo, 2020, p. 142.

443 Ivi, p. 142.

444 Lima, 2018.

445 These questions include: should the rise of more and more complex AI agents invite us to reconsider the mere notion of act as the bedrock of contemporary criminal law theory? Will we perhaps need to replace or expand or enrich the arguably obsolete notion of "voluntary bodily movement" against this new landscape? Ivi, p. 681.

446 The Model Penal Code (MPC) at § 2.01 defines criminal liability as follows "A person is not guilty of an offense unless his liability is based on conduct which includes a voluntary act or the omission to perform an act of which he is physically capable." In the "General Definitions", an act is defined as a bodily movement and Lima claims, "(whether voluntary or not), […] the act requirement is widely regarded as the most notable, or perhaps the only, exception to the rule that substantive criminal law in the United States is not regulated under constitutional law". Lima, 2018, p. 679.

447 "[T]he prevailing opinion among criminal law scholars is that an act has to be controllable by the actor and "socially relevant" – in other words, it needs to convey social meaning. Ivi, p. 680.

448 Lima, 2018, p. 680.

449 Gless, Silverman & Weigend, 2016, p. 417 analyzed above at Ch. 3.4.5.1.

do not yet seem to possess the potential for fully independent, even self-destructive decisions. In other words, no one would regard a robot's choice to change its route when stumbling upon a table as voluntary, so long as the robot is simply following an algorithm, however intricate, that dictates it to change route when encountering a physical obstacle-or to put it more simply, so long as the robot does not have the choice to keep hitting at the obstacle if it so wishes. This holds true even when this choice is the only reasonable one and in the AI agent's "benefit" of achieving its objective.[450]

The reference in the passage above to "changing routes" and to the concept of "benefit" bring up familiar concepts. As a matter of fact, both Pagallo and Sartor-Lagioia refer to these concepts when they reconstruct intent according to the BDI model.[451]

Next, Lima moves on to analyze the questions of blameworthiness and punishment. With regards to the first, she starts the discussion by noticing how *mens rea* and blameworthiness were first theorized as safeguards against state power. They restrain punishment only to those who chose to make the choice to inflict harm "conscientiously", since they had the freedom to act against the law. Actually, "criminal liability is a response reserved for those who could have risen to the occasion but chose not to".[452] She acknowledges, indeed, that this might be a shortcut to bypass discussions on "how human intent is formulated as well as any doubts about whether our free will is indeed free and our own after all".[453] Yet, this operation is not new for criminal law and the law in general. The heart of modern criminal law, according to Lima, lies in the humanity of the perpetrator, or in the "human experience". And said humanity is not possessed by AI agents, where instead it might be possessed by corporations by proxy. In other words, according to Lima the assertion that corporations are made of humans, while AI systems are not, is unsurmountable. Moreover, not only are corporations deeply intertwined with their human agents, but they are also a fiction created by them. AI systems, instead, might be said to live beyond this fictive veil, as they can, more and more, act without the involvement of humans.

With regards to punishment, Lima claims that it is a "collective mean of responding to crime directed at an agent that can understand its significance as well as its relevance to their

---

[450] Lima, 2018, p. 683.
[451] See Para. 3.3.2. and Para. 3.4.2.
[452] Lima, 2018, p. 687.
[453] Ibid.

criminal conduct",[454] and since AI systems cannot do so, the debate on applying criminal sanctions is "misplaced".[455]

Finally, Lima discusses potential models for ascribing criminal liability in the case of an AI "action" or "omission". After dealing with instrumental (mis)uses of AI systems, she focuses on negligence and recklessness. The first, she contends, represents the model that can be used most appropriately in cases where a benevolent designer or operator neglected to "take due care in order to prevent an undesirable outcome that could occur within the usual performance of the AI agent and which the programmer or the user should have foreseen".[456] The second represents the model to use – according to the different jurisdictions – in cases where "the human agent actually foresaw the outcome and decided to disregard it".[457]

Then, the author analyzes the notion of *respondeat superiror*. This concept is derived from tort law and implies the responsibility of a person who is in control of another. She argues that the analogy between the relationship master-agent and human-AI works only at a first glance when applied to criminal law. Indeed, it is true that also in cases of *respondeat superior* the *inferior* is an independent and intelligent being. Yet, criminal law demands a higher threshold for imposing liability on the "controller" and this reasoning should be extended to all the models for ascribing criminal liability:

> any potential model of ascribing liability for the human agent who is somehow involved in a crime committed by an AI agent will have to vary not only depending on circumstances, such as the sophistication of the intelligence of the AI agent or the degree of control of the human agent, but also on the type of crime committed. In other words, the threshold should be higher for serious crimes, such as killing, and could be lower for relatively minor ones, such as the destruction of an inexpensive item that belongs to a third party.[458]

---

454 Lima, 2018, p. 689.
455 Ibid.
456 Lima, 2018, p. 691.
457 Ivi, p. 692.
458 Lima, 2018, p. 693.

She then affirms that strict liability might be best considered in combination with negligence requirements, similar as it is done in the field of product liability.

In conclusion, she claims that even if humans did everything right, AI malfunction might still happen. In these cases, much as it is done in the case of a bridge collapsing, we should learn how to live with it. This, she claims, is in line with Gless et al.'s argument of considering AI systems as "exceptional risk".[459] As it is remarkably stated in the conclusions:

> Not everything can be foreseen, prevented, or contained, and in everyday life there are several instances where no one is to blame-much more be held criminally liable-for an undesirable outcome. In other words, *not everything can or should be regulated under criminal law*.[460]

Undeniably, Lima believes that criminal law is not the answer (or the "appropriate vessel") to ensure AI accountability, especially as there are "softer" State powers which might take this role, such as administrative sanctions.

The punch line of Lima's article is dedicated to a critique towards the workings of Hallevy. She claims that his theory is based on a circular argument, since it takes for granted that a concept such as *mens rea*, which was designed with humans in mind, could be fulfilled by AI systems that are not – today – human-like at all. If the opposite were true, i.e., if *mens rea* requirement could be fulfilled by non-humans, this would also "presuppose the perception of historically and empirically informed concepts such as choice, voluntariness, knowledge, and intent as simply technical terms without any inextricable grounding in the human experience".[461] We, indeed, haven't reached that point (yet).

3.4.4 Peter Asaro: A Body to Kick, but Still No Soul to Damn

Peter Asaro's catchphrase "A Body to Kick, but Still No Soul to Damn"[462] had quite the success amongst the scholars who participate to the debate object of this analysis.[463] Asaro,

---

[459] Gless, Silverman & Weigend, 2016, p. 19.
[460] Lima, 2018, p. 694 [emphasis added].
[461] Ivi, p. 696.
[462] Asaro, 2012, pp. 169-186.
[463] The work of Asaro is expressly referred to by Mulligan at p. 15, where she claims that robot punishment advances a different goal of punishment than retribution, reform or deterrence; by Abbott and Sarch, see

as a philosopher, acknowledges from the start that while "there are instances where what is legal is not necessarily morally esteemed, and what is morally required may not be legal, there is a significant overlap between what is legal and what is moral".[464] He contends that punishment is traditionally conceived as "corrective in one or more senses": through retribution one pays her debt to society; through reform one is to be reeducated so not to repeat the offense; through deterrence other people in society are deterred from committing a similar offense.

Moreover, he believes that applying criminal law to robots brings about two fundamental issues: one, criminal actions require moral agency; two, it is unclear whether a robot can be punished. The two issues, then, are deeply intertwined: only moral agents, as part of our society, can create a debt and eventually pay it, or else what we are facing is just an accident or an act of nature. Moreover, reformation entails developing or correcting moral character – or else it is just a matter of fixing a problem – and this can be done only to moral agents. What is more, only moral agents are capable of being deterred, because they are the only ones that can recognize how their actions are similar to those of other moral agents who have been punished for wrongdoing and, as a result, be deterred by it. In other words, only moral agents are capable "reflexivity of choice" and "recognition of similarity between and among moral agents".[465]

Afterwards, the author moves to analyzing the similarities and differences between imposing criminal punishment on corporations and on robots. The most obvious difference, he asserts, is that "robots do have bodies to kick, though it is not clear that kicking them would achieve the traditional goals of punishment".[466] First of all, corporations are created to make money, hence monetary sanctions would be effective because they would target a corporation's essential purpose. The same cannot be said about robots, as their purposes are not as straightforward. Second, with regards to corporal punishments, robots "do have

---

Sec. 3.3.1, at p. 345, when they refer to the consequentialist benefits which would derive from punishing AI and claim that Asaro did not make a distinction between general and special deterrence. He is quoted by Gless, Silverman & Weigend, 2016, p. 12, n.12; by Cappellini, 2018, p. 13; by F. Basile, "Intelligenza artificiale e diritto penale: quattro possibili percorsi di indagine", *DPU*, 2019, pp. 27-28; Lagioia & Sartor, 2020, p. 22.

464 Asaro, 2012, p. 169.

465 Asaro, 2012, p. 181.

466 Ivi, p. 182.

bodies to kick, though it is not clear that kicking them would achieve the traditional goals of punishment". Asaro claims that physical punishment requires something more than a body, such as desires and fears, and this is something that is not possessed by robots.

To conclude, even though it might be feasible from a technological perspective to punish a robot, this would not achieve retribution, reform or deterrence. As a matter of fact, Asaro believes that this is a "a greater hurdle to ascribing moral agency to robots directly than other hurdles, such as whether it is possible to effectively program moral decision making".[467]

3.4.5 The German Approach

The debate in German criminal legal doctrine differs from the Italian and the English-speaking fronts for two reasons. First, researchers published their work both in German and in English. Second, German speaking authors make little or no mention whatsoever in their papers to authors who are not of German provenance or who have not written in German. The reasons could be different and certainly have deeper roots.[468] One could ask herself whether this represents an instance of the traditional approach of German legal doctrine to comparative criminal law, which has been labeled by some as "self-referential" [469] and, in the words of George Fletcher, as "*selbstbewusste Provinzialität*".[470] According to some, it is (also) a matter of language, specifically of the difficulty to translate foreign legal concepts into German criminal law dogmatics.[471] Others contend that German criminal law is nothing but

---

[467] Asaro, 2012, pp. 182-183.

[468] For instance, traces of such an attitude could be found in the scholarly use of foreign law in the debate on the position of national constitutional courts in the EU. On the matter see N. Graaf, *Judicial Influencers: Scholarly use of foreign law and the convergence of German, Italian and French ideas on the position of national constitutional courts in the EU legal context*, 1989-2012, PhD Dissertation, Utrecht University Repository, 2022. For an intriguing analysis of the non-neutral nature of German (public) law libraries, see N. Graaf, "Why German Law Libraries Are Not Neutral and Why We Should Care", *LawLog Blog*, July 2019. Available at: https://lawlog.blog.wzb.eu/2019/07/25/why-german-law-libraries-are-not-neutral-and-why-we-should-care/.

[469] See *ex multis,* M. Donini, "An impossible exchange? prove di dialogo tra civil e common lawyers su legalità, morale e teoria del reato"*, Rivista Italiana di Diritto e Procedura Penale*, Fasc.1, 2017.

[470] G. Fletcher, "Deutsche Strafrechtsdogmatik aus ausländischer Sicht", in A. Eser, W. Hassemer & B. Burckhardt (Eds.), *Die deutsche Strafrechtswissenschaft vor der Jahrtausendwende,* Beck, 2000, pp. 235 ff.

[471] " *Wer die deutsche Sprache als Fremdsprache erlernt, bekommt mit dem Erlernen gleich die Terminologie der deutschen Strafrechtsdogmatik mitgeliefert. Er ist daher eher in der Lage, einen Vergleich mit den Gegebenheiten in seiner eigenen Strafrechtsdogmatik herzustellen. Für deutsche Muttersprachler ist es schwieriger, Sachgegebenheiten in einer fremden Sprache*

"provincial", even though it must be conceded that it is focused mostly on the "export" of German criminal law ("*Strafrechtsexport*"), rather than on the import of foreign legal concepts ("*Strafrechtsimport*").[472] Surely, it strikes as odd, for example, that, compared to the Italian front, neither Beck, nor Gless-Weigend-Silverman, make any reference whatsoever to Hallevy's theories nor to Pagallo's colossal work on AI and law. As it was established above, the indifference is not reciprocated by Italian scholars, who, more often than not, wink their eyes to their German colleagues.

3.4.5.1 Sabine Gless, Thomas Weigend, and Emily Silverman

This section will analyze the reflections contained in two articles, namely "Intelligente Agenten und das Strafrecht"[473] by Sabine Gless and Thomas Weigend and "If Robots Cause Harm – Who Is to Blame? Self-Driving Cars and Criminal Liability" by Sabine Gless, Thomas Weigend and Emily Silverman.[474]

In the article "Intelligente Agenten und das Strafrecht", Gless and Weigend scrutinize two issues under the lenses of German criminal law: first, the question of direct liability of

so zu erfassen, dass man sie mit der eigenen deutschen Begrifflichkeit der Strafrechtsdogmatik vergleichen kann". W. Gropp, "Deutsches Strafrechtsdenken im Europäischen Kontext" in K. Karsai (Ed.), "Strafrechtlicher Lebensschutz in Ungarn und in Deutschland. Beiträge zur Strafrechtsvergleichung", *Stiftung Elemér Pólay*, 2007, p. 18.

472 "*Immer mehr Strafrechtslehrstühle werden bewusst mit internationaler oder europäischer Ausrichtung versehen; die deutsche Lehrbuchliteratur zum internationalen und europäischen Strafrecht ist sehr reichhaltig; auch in der übrigen Literatur sind strafrechtsvergleichende, europäisch- und internationalstrafrechtliche Beiträge mittlerweile häufig. Trotzdem ist der Kritik zuzugeben, dass die deutsche Strafrechtswissenschaft, soweit sie ins Ausland bzw. auf die europäische oder internationale Ebene zu wirken unternimmt, eher an einem 'Strafrechtsexport' interessiert erscheint, also daran, deutsches Strafrechtsdenken im Ausland bzw. in europäischen oder internationalen Gremien zu verbreiten, als daran, vom ausländischen, europäischen oder internationalen Strafrecht zu lernen und gleichsam einen 'Strafrechtsimport' zu betreiben, der keineswegs zwingend auf ein punitiveres Strafrecht hinauslaufen müsste, sondern selbstverständlich auch alternative Modelle mit umfassen könnte. Diese mangelnde „Importbereitschaft" kann sich nachteilig auswirken, insbesondere wenn große ausländische Namen mit richtungweisenden Werken weithin unbekannt bleiben und die deutsche Diskussion nicht bereich*.". J. Vogel, "Strafrecht und Strafrechtswissenschaft im internationalen und europäischen Rechtsraum", *ZIS*, No.1, 2012, p. 27. For a critical perspective, see B. Schünemann, "Über Strafrecht im demokratischen Rechtsstaat, das unverzichtbare Rationalitätsniveau seiner Dogmatik und die vorgeblich progressive Rückschrittspropaganda", *ZIS*, Vol. 10, 2016, pp. 654-671.

473 Gless & Weigend, 2014, pp. 561–591.

474 Gless, Silverman & Weigend, 2016, pp. 412-436.

AI system, second, the question of liability of "*[d]er Mensch hinter dem Intelligenten Agenten*",[475] i.e., the human behind the machine.

With regards to the first aspect, they initially focus on the legal personhood of AI systems. They contend that AI systems today are not capable of being accountable for their own actions, since ultimately they always follow the instructions encoded in their programs, and, consequently, they cannot be regarded as possessing personhood under German criminal law.[476] Then, they focus on *actus reus* (*Handlungsfähigkeit*). Interestingly so, they acknowledge that the issue of whether an act of an AI system can be considered as relevant under criminal law is strictly dependent on which theory of conduct one adopts:

> According to the traditional "causalistic" theory, any voluntary bodily movement is sufficient to constitute a criminally relevant act; its more modern "social" variant requires, in addition, some kind of societal reference of the bodily movement.[477]

From a causalistic perspective, it follows, it is easy to conclude that any movement of AI systems that operate directly in the physical sphere, such as robots, constitutes an "act", since the threshold of volition required to consider such movement as relevant is very low. If one considers, instead, a finalistic standpoint, the act is relevant under criminal law only as long as it is "the expression of a purposeful will of the actor".[478] In this perspective, according to Gless and Weigend, AI systems are not capable of committing an act "willfully": even if it is true that they can set their own goals and determine autonomously how to reach these goals, this does not mean that they are aware of the relevance of their actions. The only thing that they are aware of is that their conduct will lead them closer to obtaining the programmed goal. The authors conclude this reflection on *Handlungsfähigkeit* by regarding it as a matter of adopting a "thick" or a "thin" definition of act[479]:

---

475 Gless & Weigend, 2014, p. 579.

476 Ivi, p. 570.

477 "*Für die überkommene 'kausalistische' Theorie genügt jede gewillkürte Körperbewegung als strafrechtlich relevante Handlung; ihre modernere 'soziale' Variante verlangt darüber hinaus einen irgendwie gearteten Sozialbezug der Körperbewegung*". Gless & Weigend, 2014, p. 571.

478 "*Nur wenn und weil die menschliche Handlung Ausdruck eines zielgerichteten Willens des Handelnden ist, kann sie strafrechtlich relevant sein*". Ivi, p. 572.

479 Gless, Silverman & Weigend, 2016, p. 420.

> In a "causalistic", merely extrinsic view, which defines every "arbitrary bodily movement" as an action, they are to be regarded as agents by all means. The more the concept of action is substantively loaded, the more one reads into it a self-conscious determination of goals, the less Intelligent Agents can meet the requirements of the ability to act.[480]

When discussing culpability (*Schuldfähigkeit*), the authors claim that according to a (German) constitutional oriented interpretation, criminal liability is meant as strictly personal, pursuant to the constitutional principle of blameworthiness. Personal criminal liability presupposes that a person is free, responsible, and capable of moral self-determination. Therefore, it entails that she is able to decide in favor of the right and against the wrong, of arranging her behavior according to the standards of the legal ought, and of avoiding what is legally forbidden as soon as she has attained moral maturity.[481] According to the authors, even the most sophisticated AI systems are not capable, today, of such moral self-determination, hence they cannot be deemed culpable.[482]

The authors then contend that recent neuroscientific evidence which questions the concept of free will as we know is irrelevant in relation to the *mens rea* of AI systems. Indeed, for a time now German authors have argued in favor of a functional conception of criminal culpability, which entails that "the attribution of criminal culpability to a person presupposes that the person had the capacity to put into question the validity of a legal norm".[483] That being the case, they proclaim that culpability is based on whether an actor can engage in a normative discourse. In order to participate in such a dialogue, the agent needs to be able to

---

480 "*Bei 'kausalistischer', bloß äußerlicher Betrachtung, die jede 'willkürliche Körperbewegung' als Handlung definiert, sind sie durchaus als Handelnde anzusehen. Je stärker man den Begriff der Handlung substantiell auflädt, je mehr an selbstbewusster Zielbestimmung man in ihn hineinliest, desto weniger können Intelligente Agenten den Voraussetzungen der Handlungsfähigkeit genügen*", Gless & Weigend, 2014, p. 572.

481 "*Der innere Grund des Schuldvorwurfes liegt darin, daß der Mensch auf freie, verantwortliche, sittliche Selbstbestimmung angelegt und deshalb befähigt ist, sich für das Recht und gegen das Unrecht zu entscheiden, sein Verhalten nach den Normen des rechtlichen Sollens einzurichten und das rechtlich Verbotene zu vermeiden, sobald er die sittliche Reife erlangt hat* (…)". Bundesgerichtshof (BGH), 18.03.1952, GSSt 2/51, pp.16-17.

482 Gless, Silverman & Weigend, 2016, p.421.

483 Ibid.

self-reflect about his past actions and to evaluate them against a moral reference system, i.e., to have a good or bad conscience.[484]

This leads back to the initial conclusion: even though there are efforts today to program AI systems which can make decisions according to moral standards (based on "merits" and "demerits"), current AI systems are far from being moral agents. Hence, they cannot be blamed. This could change in the future in case AI systems capable of moral evaluations were created. Admittedly, "[t]he final step would be to make robots capable of understanding punishment, that is, to teach them to associate certain changes in their environment with the wrongfulness of their prior acts".[485]

Let us move now to the analysis of the second part of the authors' work, which deals with criminal liability of the "human behind the machine". The focus is on negligence offenses. According to German criminal legal doctrine, negligence is based on two elements: (a) foreseeability and (b) breach of a duty of care.

With regards to (a), the authors acknowledge that certain AI systems have a sort of "pre-programmed" unpredictability, meaning that "the operator cannot reduce to zero the possibility that robots may cause harm to others".[486] From this assumption, they theorize two mutually exclusive conclusions (regarding the programmer's liability for negligence):

> It could be argued that he cannot be held responsible because the machine is acting "on its own"; alternatively, it could be claimed that he can foresee any and all harm that robots might cause and therefore should face de facto strict liability for the results of the robots' acts.[487]

The first argument is discarded, as it would lead to a gap in the protection of individuals' interests. Admittedly, it would be as if the manager of a zoo, who released a tiger

---

[484] Gless & Weigend, 2014, p. 575.
[485] Gless, Silverman & Weigend, 2016, p.424,
[486] Ivi, p. 426; Gless & Weigend, 2014, p. 581.
[487] Gless, Silverman & Weigend, 2016, p.426.

from its cage, remarked the unpredictable nature of the wild animal to the innocent bystander who had just been attacked and bitten by the very same tiger.[488]

If the second argument stands, then this entails that

> a person who can foresee that his action might harm interests protected by criminal law (such as the life and health of other persons) is obliged to refrain from that action. Hence, if the zoo manager can foresee that the tiger, if set free, will harm human beings, he must refrain from releasing the tiger from its cage.[489]

With regards to (b), the authors claim that German case law has developed strict rules which would be applicable to AI systems with regards to due diligence, especially in the field of product liability. Criminal product liability is based on the fact that "a perfectly legal act – [for example] the marketing of a self-driving car in accordance with the current state of knowledge and technology – may trigger criminal liability for omission",[490] the omission being that the producer did not ensure that the product adhered to certain standards of safety. Naturally, such a strict application of the duty of care will hinder technological progress and therefore diminish, if not erase the benefits that our society is enjoying due to the innovation surge. For these reasons, the authors argue that such a comprehensive criminal liability of the programmer, which includes any foreseeable damage, should be excluded for the time being.

The crucial question, then, is the concept of permissible (innovation) risk: where should the line be drawn? Even in cases where society could accept innovation risks, such as with AVs, the producer would still have to make sure to reduce risks to the minimum, by complying with other duties such as information duty towards the users and the processes of monitoring/recalling the products.[491]

Assuming that conditions (a) and (b) are fulfilled, the authors then discuss whether the harmful event can be causally attributed to the programmer or not. The first option would

---

488 *"[…] er kann dies ebenso wenig wie der Zoodirektor, der einen Tiger in die Freiheit entlässt und gegenüber einem Passanten, der von dem Tiger angefallen und gebissen wird, auf die unberechenbare Natur des wilden Tieres verweist*". Gless & Weigend, 2014, pp. 581-582.

489 Gless, Silverman & Weigend, 2016, p. 427.

490 Ivi, p. 428.

491 It is referred to as a "sleeping obligation" (*schlafenden Ingerenz)* by Gless & Weigend, 2014, p. 585.

be to regard AI systems as part of the average risks of life, same as lightning or falling trees.[492] Yet, this option is not feasible, since nowadays advanced AI systems are considered as exceptional risks. This could change in the future, as AI systems permeate our lives more and the technology progresses. The authors subsequently suggest reducing the duties of operators of (some) robots "to employing the best knowledge and technology available in manufacturing, programming, testing and monitoring them".[493]

Conclusively, Gless and Weigend affirm that society will have to accept a degree of (residual) risk for the sake of a greater social benefit: instead of imposing a total ban of AI systems because they pose uncontrollable risks, the option would be to "burden "society" with the dangers that cannot be reliably controlled by programming and responsible use. This would happen by waiving criminal liability for negligence, i.e., by defining the injured persons as victims of non-human behavior.[494]

3.4.5.2 Susanne Beck

Beck claims that the debate on how the law should deal with the risks of robots resembles other debates on the legal handling of risks (the "risk society"). She focuses, then, on the criminal regulation of risks, specifically on negligence. Her reflections on the interaction between legal and non-legal standards in determining the boundaries of negligence are indeed noteworthy.

Negligence, she notices, is made of two requirements: reasonable care and foreseeability of the damage. With regards to the first requirement, i.e., reasonable care, usually the standard of care is determined by how society expects that a reasonable person would act in a situation. Sometimes, the standard is set by non-legal sources and standardizing institutions (e.g. ISO). Thus, in the field of robotics (and AI systems in general, one might add) the development of these standards has not been completed yet. What is more, Beck argues, one must not forget that technical standards serve a different purpose

---

[492] Gless, Silverman & Weigend, 2016, p. 433.

[493] Ivi, p. 434.

[494] "*Will man sich nicht wegen der letztlich nicht beherrschbaren Risiken auf ein Totalverbot Intelligenter Agenten verständigen, so dürfte keine andere Möglichkeit bestehen, als die nicht verlässlich durch Programmierung und verantwortlichen Einsatz steuerbaren Gefahren der „Gesellschaft" aufzubürden, indem man auf eine strafrechtliche Fahrlässigkeitsverantwortlichkeit verzichtet, also die geschädigten Personen als Opfer nicht-menschlichen Verhaltens definiert*". Gless & Weigend, 2014, p. 590.

than criminal law: they aim at minimizing risks and preventing danger while fostering economic advantages. Criminal law is not "simply an accessory to the regulations of non-governmental groups",[495] it has also to consider whether the action is socially inadequate, i.e., whether it violates social and moral rules. Nevertheless, our society needs subsystems (such as the technology one) and accepts them, which entails that "[i]t would be inconsistent to rely on these systems on one side and not to accept their specific norms which regulate these subsystems and the interests of its parties on the other. Thus, the inclusion of economic interests in standardising procedures does not necessarily lead to their irrelevance for criminal law".[496] In point of fact, Beck does not see a point in time when these norms would become irrelevant for criminal law. This, she believes, is especially relevant in the field of robotics: first, standardizing institutions are very active in the robotics field right now, hence "it seems, from a legal perspective, important to analyse these activities and retie them with legal evaluation. One might even have to consider interaction with the standardising institutions to secure plausible normative premises and processes".[497]

Second, and most interestingly,

> Most researchers and producers are convinced to have acted legally when complying with the existing standards, even if they are somehow vague, not covering all relevant (dangerous) aspects of their activities and normatively questionable. It is necessary to discuss how to connect this strong conviction, supported not just by the official impression of standardising institutions but by the general custom in the actors community, with negligence liability; it might be worth to consider its relevance for the subjective aspects of negligence (guilt). The (potential) "sense of right and wrong" is part of liability for negligence as well. Unavoidable mistake in the lawfulness of the action can therefore lead to negation of negligence.[498]

---

[495] S. Beck, "Intelligent Agents and Criminal Law—Negligence, Diffusion of Liability and electronic personhood", in E. Hilgendorf and J. Fedle (Eds.), Vol. 86, 2016, p. 139.
[496] Ibid.
[497] Beck, 2016, p. 139.
[498] Ibid.

This would be the case particularly for those who are not part of the producing chain of the AI system, who would therefore be "surrounded by a community in which everyone is convinced that fulfilling the requirements of standards is sufficient to act lawfully".[499]

With regards to the second requirement, i.e., foreseeability of damage, Beck notices that the more the AI system has the faculty to behave autonomously, the more "it can be – generally – foreseen during the research phase that it may, later on, bring harm to humans".[500] Yet, in this sense foreseeability would be related only to a *general* risk of harm, where instead "the specific conditions and situations become more and more unforeseeable".[501] Consequently, Beck states that via analyzing negligent crimes connected to AI harm one can reflect on which degree of *specificity* is required for the fulfilment of the foreseeability-component of negligence. Specifically, one must understand whether the mere possibility of "violating humans as such"[502] would be sufficient.

Having said this, Becks shifts her attention on how the aforementioned traditional notion of negligence could be adapted to the risk society. She contends that, rather than focusing on foreseeability, or on external regulations which would develop the required standard of care, we should concentrate respectively on the social adequacy of the action and on the legal construction of admissible risk. In other words, one should focus on how to "negotiate in each area of life if and under which conditions the usage of robots is regarded as such 'admissible risk' and if one does act in the adequate framework, one cannot be responsible for the consequences hereof".[503] Conclusively, she argues in favor of restricting the use of criminal law in the field of technology development. This makes her approach to the subject both original and cogent.

3.4.5.3 Gerhard Seher

Gerhard Seher, as most of the other authors discussed here, develops his paper following the building blocks of criminal law.[504] Therefore, he starts by analyzing whether AI systems

[499] Beck, 2016, p. 139.
[500] Ibid.
[501] Ibid.
[502] Ibid.
[503] S. Beck, "Robotics and Criminal Law. Negligence, Diffusion of Liability and Electronic Personhood", in Hilgendorf & Feldle, 2018, p. 53.
[504] Seher, 2016.

can "act" and can be considered as "agents". Seher believes that the chore of a criminal act is highly normative: criminal law shall concern itself only with those actions that show at least potential norm understanding. Surely, in the history of criminal legal doctrine the meaning of act varied and so did the demand of a "free will" component in it, similarly to a sine curve with its high and low oscillations.[505] Even if we were to accept that, as humans, we do not possess free will, our "determined" brain is still able to understand and reflect legal commands. This means that, even from a strictly deterministic point of view, we are still influenced by legal norms and, therefore, our actions still amount as such from the perspective of criminal law.

Shifting to AI agents, Seher believes that computers are not capable, as of now, to perceive norms as such by themselves.[506] The mediation of a human programmer is always needed. Seher then argues that the situation might be different when it comes to autonomous vehicles: one could argue that they have indeed the capability of understanding symbolized norms such as traffic signs and traffic light signals since they represent "schematized, symbolized norms, which are directly perceptible and implementable for a computer".[507]

Moreover, Seher contends that the "*Rechtlich relevante Handlung*" can be defined only as the action of a person that is responsive to the law and this has nothing to do with determinism or free will. This person needs to be able to directly understand the meaning of the norm and to include it into her decision-making process. When claiming so, Seher takes distance from Gless und Weigend: he claims that they take a "cautiously different" approach which is based on a "norm-theoretically open, descriptive concept of action".[508]

At this point of the discussion, Seher shifts his attention to the responsibility of *human* actors involved with AI systems. He claims that since he already made a strong case against AI systems being capable of committing an "act", the matter does not deserve further attention. If one considers causality, then, he argues that the most interesting study cases are

---

505 Ivi, p. 48.

506 "*Computer sind – soweit ich informiert bin – nicht in der Lage, von sich aus Normen als solche wahrzunehmen"*. Seher, 2016, p. 50.

507 "*Das mag man bei selbstfahrenden Autos möglicherweise anders sehen, soweit sie in der Lage sind, Verkehrsschilder und Ampelsignale richtig zu „lesen", denn hier handelt es sich um schematisierte, symbolisierte Normen, die für einen Computer unmittelbar wahrnehmbar und umsetzbar sind*". Ibid.

508 "*Das sehen Gless/Weigend […] vorsichtig anders, weil sie bereit sind, von einem normtheoretisch offenen, deskriptiven Handlungsbegriff auszugehen*", Seher, 2016, p. 51, n. 11.

those where a correctly programmed AI system makes a decision which results in damage. In fact, as Seher points out, the sole fact of creating an AI system creates the abstract risk that it might damage a legal good, similarly as it already happens with cars and airplanes. If we were to follow the standard rule on attribution, i.e., the operation through which we select the legally relevant cause of a harmful event, then the operators would always be liable for damage cause by the AI systems they created. However, Seher argues that there are certain conditions which would stop the attribution process: to begin with, if the misconduct of the AI agent is the result of an atypical causal process according to general life experience. To continue, when the conduct of the AI systems falls in the scope of a permitted risk.

Next, Seher spends a short paragraph on *mens rea* of humans and in this he differs from most of the authors analyzed in this Chapter, who instead believe that it is a pivotal issue. This represents a direct reflection of the German tripartite conception of criminal liability. Notably, he argues that, when it comes to human operators and negligence, the general foreseeability of a harmful event is not sufficient. What is needed is foreseeability tailored to the specific error that caused to the concrete harmful event.

Seher, then, focuses on blameworthiness, defined – following what he claims to be a "unanimous opinion" – as the judgment from a legal community which deems a certain action as reproachable. As a matter of fact, those in favor of a functionalist theory of criminal law see the criminal accusation as a "branding operation". It is a tool to distance ourselves from the criminal act and the offender.[509] Based on this approach, Seher argues that AI systems should not be blamed: our society does not consider their behavior as an attack towards the validity of the norm.

Finally, with regards to punishment, Seher claims that assuming that we could punish AI systems at all, it is uncertain whether inflicting this punishment would fulfill any purpose. First, it would not provide retribution to the victims. Second, general prevention would not be achieved since – as it was stated above – AI misbehavior would not be perceived as the violation of a norm. Third, even if we could program an AI system to perceive a punishment for violating a norm as a failure, and to learn from it, it still would not be aware that it is being punished. AI systems are not capable of perceiving the reprimand of a criminal sanction, hence their behavior cannot be guided by it.

---

[509] Ivi, p. 55.

Indeed:

> Punishment is a concept that belongs in the context of normative understanding. However, no communicative understanding in a normative language occurs between "intelligent agents" and humans. These agents can only be addressed via binary digital coding. Such systems can be reprogrammed or shut down, but not punished.[510]

3.5 CONCLUSIONS

The purpose of this Chapter was to provide an across-the-board analysis of the current debate on criminal liability connected to AI systems. The word "connected" is important in this sense: this Chapter covered not only the ideas of authors on how to impute liability *directly* on the artificial agent, but also on how to impute it on those (humans) who surround it.

The debate was divided into three streams: expansionists, moderates, and skeptics. Expansionists are looked at by the rest of the scholars with surprise and, at times, disapproval. They are a rare breed: this research, indeed, found only a few of them. On the opposite side, skeptics take up the cudgels for criminal law's traditional principles and defend them from the disruptive effects of AI systems. The moderate stream is multifaceted since it comprises of authors with different approaches and opinions.

What is the significance of this Chapter? Surely, providing a *vision d'ensemble* of such a rich and various debate while keeping an international breath is valuable *per se*. Admittedly, this operation is paramount of what should be regarded as the *credo* of all the scholars interested in the newborn realm of AI law: No Law is an Island.[511] But there is more. The

---

[510] "*Strafe ist ein Begriff, der in den Kontext normativer Verständigung gehört. Zwischen „intelligenten Agenten"und Menschen findet aber keine kommunikative Verständigung in einer normativen Sprache statt. Ansprechbar sind diese Agenten nur über eine binäre digitale Codierung. Solche Systeme kann man umprogrammieren oder stilllegen, aber nicht bestrafen*", Seher, 2016, pp. 59-60.

[511] The expression is borrowed from a passage of the famous 1624 Meditation XVII by John Donne, which reads: "No man is an island, entire of itself; every man is a piece of the continent, a part of the main. If a clod be washed away by the sea, Europe is the less, as well as if a promontory were, as well as if a manor of thy friend's or of thine own were: any man's death diminishes me, because I am involved in mankind, and therefore never send to know for whom the bells tolls; it tolls for thee". J. Donne,

investigation started by asking whether the international legal debate of on the interaction between AI systems and criminal liability was running on common tracks. Indeed, regardless of which side they were on, most scholars seemed to follow a similar structure when discussing this topic: first they define AI, then they review *actus reus* and *mens rea* requirements, finally they discuss AI-punishment. As mentioned above, the debate circles around two macro-areas: on one side, the direct liability of AI agents, on the other, the liability of humans in the loop. The two areas are interconnected.

Finally, it possible to formulate a set of intermediate conclusions. Upon conducting this literature review, it was possible to identify a set of recurring questions and topics discussed by scholars:

I Are robots/AI systems capable of moral reasoning?
II What is the relevance of moral reasoning when ascribing criminal liability?
III Can an AI system be culpable?
IV Why should we punish AI systems and how should we punish AI systems?
V Can an AI system "act"?
VI Permitted risk and dangerous activities;
VII Should human agents be liable for the AI's misbehavior?
VIII The creation of new negligence crimes and the application of existing negligence offenses so to humans-in-the-loop (how should foreseeability and reasonable care be defined?);
IX What is the relationship between corporate criminal liability and AI criminal liability?
X Are there more feasible alternatives to applying criminal law (such as torts law or administrative sanctions)?

These questions will influence the analysis conducted in this research. The next chapters will necessarily address them, either directly or indirectly. Specifically, questions I to IV will be touched upon in chapters 4 and 5; question V will be addressed at Paragraph 6.3.; questions VI to VIII will be tackled at Paragraphs 6.2.3.1 to 6.2.3.4; question IX will be part of Paragraph 6.4. and question X will be discussed throughout Chapter 8.

"Meditation XVII. Nunc Lento Sonitu Dicunt, Morieris", in A. Raspa (Ed.), *Devotions Upon Emergent Occasions*, Oxford University Press, 1987.

Is there a gap to be filled, then, by this very research? It is argued here that the answer is positive. It is possible to identify the following gaps and weaknesses in the discussion conducted by scholars:

I Civil criminal legal scholars discard Hallevy's theories quite effortlessly. They share a defensive approach of the principle of *personal* criminal liability, as in belonging to a human *persona*. This "as sure as eggs is eggs" approach, at times, undermines the strength of their legal arguments;

II No or little attempts at providing a definition of AI and of AI agent applicable to criminal law, no discussion of the impact of the lack of a definition on the matter;

III No or little attempts at addressing the preliminary issue of considering AI as a subject of criminal law (i.e., criminal capacity of AI);

IV Most legal scholars neglect discussing whether an AI system can act, think or want. The topic is mostly addressed by philosophers and legal philosophers. For example, Abbott and Sarch claim specifically "[w]e will not attempt to articulate the non-functional differences between human and algorithmic reasoning, a subject which has fascinated and confounded computer scientists since the 1950s. […] Functionally, AI and people can exhibit similar patterns of behavior and information processing, regardless of whether machines "think" or understand what they do";[512]

V No or little discussion of whether conduct of AI systems could be excused or justified by a defense;

VI Focusing on the liability of humans in the loop, there seems to be confusion on identifying precisely who would be the responsible human agent. Authors mention different subjects (such as robot trainers, robot owners, robot users, operators, programmers, designers, users), yet their roles nor their duties are specified. For example, no author specifies that the liability of the programmers could be caused by a wrongful selection of the training data for the algorithm;

[512] Abbott & Sarch, 2019, n. 53, p. 333.

VII. While some authors discuss the notion of act, almost no attention is given on how AI behavior interferes with theories of causality.

One of the purposes of this research is to grapple with these gaps and advance the discussion on the matter. At Para 8.3, I will evaluate whether this operation was successful.

4 ASCRIPTION

The only part of conduct of any one, for which he is amenable to society, is that which concerns others.
In the part, which merely concerns himself, his independence is, of right, absolute.
Over himself, over his own body and mind, the individual is sovereign.

John Stuart Mill, Essay on Liberty, John W Parker and Son (1859)

4.1. Introduction**.** On Paperclips, Planes and AI **– 4.2.** Methodology and Structure of Chapter

4.1 INTRODUCTION. ON PAPERCLIPS, PLANES, AND AI

In 2003,[513] Nick Bostrom, an Oxford-based philosopher, conceived the "Paperclip Maximizer" ("PCM"),[514] a thought experiment where he hypothesized the emergence of an AI system characterized by general superintelligence,[515] whose only goal would be manufacturing as many paperclips as possible, while opposing resistance to anybody or anything that would attempt to modify its goal. As a consequence, the superintelligence would start transforming all of earth (and then the space) into paperclip manufacturing facilities. This thought experiment is used to illustrate how an AI, tasked with a worthless and innocuous utility function,[516] such as producing the biggest number of paperclips in the universe, could "as a side-effect destroy us by consuming resources essential to our

[513] N. Bostrom, "Ethical Issues in Advanced Artificial Intelligence", 2003. Available at: https://nickbostrom.com/ethics/ai.

[514] The PCM Experiment is mentioned in M. E. Diamantis, R. Cochran & M. Dam, "AI and the Law: Can Legal Systems Help Us Maximize Paperclips while Minimizing Deaths?", 2022. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4177378.

[515] Superintelligence is defined as "any intellect that is vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills". N. Bostrom, "How Long Before Superintelligence?", *International Journal of Futures Studies*, Vol. 2, 1998.

[516] In a utility function numerical values ("utilities") are assigned to choices based on the satisfaction that can be obtained from the choice (the higher the utility value, the higher the satisfaction). Utility functions are used to analyze human behavior in rational choice theory. See: https://www.lesswrong.com/tag/utility-functions.

survival",[517] leading to human extinction. Simply put, in a (perhaps) near future[518] we might face a doomsday scenario: an AI system will transform us all into paperclips.[519]

Even though the PCM apocalypse scenario is "just" a philosophical thought experiment, the last five years marked the start of collections of real-life AIs "going wrong" scenarios. One example is Awful AI, a website which contains a list of "*current scary usages of AI*"[520] such as a DNN designed to "detect sexual orientation from facial images" (also referred to as "an AI-based gay radar"),[521] or Tay, the Microsoft chatbot which engaged in racist and anti-Semitic messages on Twitter (e.g., "Hitler was right") after only 24 hours from its 'birth'.[522] Other examples of real life AI incidents databases are the AI Algorithmic and Automation Incidents & Controversies (AIAAIC) Repository[523] and the AI Incident Database (AIID).[524] These databases validate this study's assumption: the idea of AI systems

---

[517] See https://www.lesswrong.com/tag/paperclip-maximizer.

[518] Unfortunately for us humans, a number of scientists argue that it is likely that AI systems could lead to an existential catastrophe. See M. K. Cohen, M. Hutter & M. A. Osborne, "Advanced artificial agents intervene in the provision of reward", *AI Magazine*, Vol. 43, 2022.

[519] In 2017 Frank Lantz, a game designer of the NYU Game center, created "Universal Paperclips". Universal Paperclips is click game where one can play the role of an AI system producing paperclips. The game ends with the destruction of the world. The game is available here: https://www.decisionproblem.com/paperclips/index2.html. For a comment, see: A. Rogers, "The Way the World Ends: Not with a Bang But a Paperclip", *Wired*, 21 October 2017. Available at https://www.wired.com/story/the-way-the-world-ends-not-with-a-bang-but-a-paperclip/; N. Jahromi, "The Unexpected Philosophical Depths of the Clicker Game Universal Paperclips", *The New Yorker online*, 28 March 2019. Available at: https://www.newyorker.com/culture/culture-desk/the-unexpected-philosophical-depths-of-the-clicker-game-universal-paperclips.

[520] D. Dao et al., "Awful AI - 2021 Edition". Available at https://github.com/daviddao/awful-ai [emphasis added]. See also "ResponsibleAI", available at: https://romanlutz.github.io/ResponsibleAI/.

[521] Y. Wang & M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images", *Journal of Personality and Social Psychology*, Vol. 114, Issue 2, 2018, pp. 246–257.

[522] E. Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter", *The Guardian*, 24 March 2016. Available at: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

[523] AIAAIC Repository. Available at: https://docs.google.com/spreadsheets/d/1Bn55B4xz21-_Rgdr8BBb2lt0n_4rzLGxFADMlVW0PYI/edit#gid=1051812323.

[524] The AIID indexes more than 1,000 publicly available incident reports (i.e., a mixture of documents from the popular, trade, and academic press). S. McGregor, "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database", *Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21). Virtual Conference*, 2021, p. 2. Available at https://incidentdatabase.ai/?lang=en.

harming people and of consequently applying criminal law is not a matter of science fiction anymore. It is just a matter of time.

Incident databases are commonly used in the field of aviation: one usually distinguishes between "accidents", i.e., "cases where substantial damage or loss of life occurs", and "incidents", i.e., "cases where the risk of an accident substantially increases".[525] Focusing on an example, "when a small fire is quickly extinguished in a cockpit it is an 'incident' but if the fire burns crew members in the course of being extinguished it is an 'accident'".[526] Incidents databases are a form of continue self-examination, which admittedly make air travel "one of the safest forms of travel" ever invented.[527]

The criminalization of aviation, i.e., the prosecution of pilots, aviation controllers and mechanics,[528] together with airline companies and manufacturers, is indeed outside of the scope of this thesis. Yet, it presents interesting similarities with the object of the present research.[529] The reasons are manifold. First of all, aviation is a highly automated field which is characterized by complex human-machine interactions. These factors lead to difficulties in ascribing (criminal) liability to human individuals when accidents happen. Moreover, "[p]rosecution of pilots, controllers and mechanics is often based on general hazard statutes that have evolved from road traffic laws which criminalize the reckless endangerment of people or property".[530] It follows that criminal legal systems have already dealt with issues of criminal liability arising from technological failures.

Furthermore, aviation is an environment where practices of "greying *[the content]* of the Black Box"[531] have already been implemented. For example, one way aviation addresses accountability issues is through technical investigations[532] which are "non-punitive in nature" and have the sole scope of "identifying all the circumstances that led to the accident, in order

---

525 Ivi, p. 1.

526 McGregor, 2021, p. 2.

527 Ibid.

528 S. Dekker, "*Pilots, Controllers and Mechanics on Trial: Cases, Concerns and Countermeasures*", *International Journal of Applied Aviation Studies*, Vol. 10, No. 1.

529 M. C. Elish & T. Hwang, "Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation*", Comparative Studies in Intelligent Systems – Working Paper #1,* Vol. 2, 2015.

530 Dekker, 2010, p. 32.

531 I.e., understanding the contents of the black box through technical investigations.

532 International Civil Aviation Organization (ICAO), "Aircraft Accident and Incident Investigation", Annex 13 to the Convention on International Civil Aviation.

to facilitate safety recommendations to prevent similar accidents in the future".[533] Accident investigations are governed by a "non-blameworthy philosophy"[534] and are conducted separately from legal investigations, which instead are aimed at finding a culprit.[535]

The two fields (i.e., AI and aviation) are now converging, with AI based technologies currently being deployed in aviation and air traffic management (ATM) for multiple purposes, such as traffic prediction, forecast, and modelling, automated flight plan correction, and fuel optimization.[536] It is predicted that the impact of ML on aviation will be tremendous. It will lead to the development of autonomous flight (the "holy grail" of AI aviation)[537] and, more generally, to a revolution of the relationship between pilots and systems. AI systems will assist the crew in taking critical decisions, such as those regarding unpredictable situations (e.g., wind gusts or engine failures) and obstacles (e.g., birds).[538] For example, in 2020 Airbus concluded its Autonomous Taxi, Take-Off and Landing (ATTOL) project: for the first time a commercial aircraft autonomously concluded the ATTOL operation using an AI system performing automatic vision.[539]

When it comes to accountability in aviation, scholars have affirmed that

> Regulators, in addition to the engineers and managers of aviation systems, have created a schizophrenic dynamic in which automation is seen as safer and superior in most instances, unless something goes wrong, at which point humans are regarded as safer and superior. Unfortunately, creating this kind of role for humans, who must jump into an emergency situation at the last minute, is something humans do not do well.[540]

---

533 S. Michaelides-Mateou & A. Mateou, *Flying in the Face of Criminalization. The Safety Implications of Prosecuting Aviation Professionals for Accidents,* Routledge, 2010, p. 2.
534 Ivi, p. 162.
535 The authors argue that more regulation should be adopted regarding this exchange of information. Michaelides-Mateou & Mateou, 2010, p. 99.
536 EUROCONTROL, European Aviation/ATM AI High Level Group (EEAI HLG), "Fly AI report", 2020, p. 55. Avaialble at: https://www.eurocontrol.int/publication/fly-ai-report.
537 Ivi, p. 7.
538 EUROCONTROL, 2020, p. 55.
538 Ivi, p. 7.
539 Airbus, "Airbus concludes ATTOL with fully autonomous flight tests", 29 June 2020. Available at: https://www.airbus.com/en/newsroom/press-releases/2020-06-airbus-concludes-attol-with-fully-autonomous-flight-tests.
540 Elish & Hwang, 2015, p. 12.

This trend has become even more tangible with the influence of AI systems. It has been argued that the introduction of such technologies has "wrestled control away from airline pilots",[541] while they are still deemed morally responsible for their outcomes (it is the so-called "moral crumple zone").[542] Pilots are nevertheless treated as "liability sponges, absorbing in one concentrated place liability for actions which are in fact distributed throughout the system".[543]

As it will be analyzed in Chapter 7, there seems to be a similar inclination when it comes to European regulation of AI systems. Thus, as of today, no AI-specific criminal law or standards have been adopted and the discussion on how to adapt what is already in place is moving slowly. It is likely that courts and regulators approaching the field in the future will indeed look at aviation and connected liability issues. The only (premature) exceptions to this trend will be analyzed in Chapter 7, can be found mostly the field of autonomous driving. Autonomous driving shares similarities with aviation such as the delegation of control to a (semi)automated system, thus, it is characterized by much fewer expert users in comparison to pilots and aviation controllers.

4.2 THEORIES OF CRIMINALIZATION

This research will unavoidably be concerned with different criminalization theories. It might approach them directly or indirectly. Hence, it is relevant at this point to briefly introduce the topic.

Criminal law is a "multi-functional" tool.[544] The choices made by legislators when defining the latitude of the special part of the Criminal Code, i.e., "decisions about offense

---

[541] W. D. Holford, "An Ethical Inquiry of the Effect of Cockpit Automation on the Responsibilities of Airline Pilots: Dissonance or Meaningful Control?", *Journal of Business Ethics,* 2022, p. 142.

[542] It is defined as the "misattribution of responsibility for an action to a human actor who has limited control over the actions of an automated system, a recognized phenomenon within human factor studies". Ivi, p. 146.

[543] Elish & Hwang, 2015, p. 15.

[544] A. P. Simester, *Fundamentals of Criminal Law: Responsibility, Culpability, and Wrongdoing*, Oxford University Press, 2021, p.3.

descriptions",[545] are a mirror of a "political community's attitude toward citizens".[546] As such, "apart from waging war, no decision made by the state is more significant than its judgment about what conduct should be proscribed and how severely to punish it".[547] It follows that legislators will have to face similar questions in the future when dealing with cases of AI-related harm and with the criminalization of AI-conduct.

Criminalization theories aim to answer the question of "what kind of conduct should be declared criminal".[548] Amongst the most prominent criminalization theories, it possible to identify the legal goods doctrine ("*Rechtsgüterlehre*"), which entails that criminal law should focus on punishing conducts to protect legal goods;[549] and the harm principle, which entails that criminal law should punish conducts with the purpose of preventing harm to others.[550]

According to some, criminalization should occur to protect the *rights* of others.[551] According to others, criminal law should protect *moral values* (following the so-called "legal moralism" doctrine). Indeed, the different "reasons for making a form of conduct an offence are also likely to be somehow related to the reasons for considering it as wrong".[552] There are different variations of legal moralism.[553] Some, including H. L. A. Hart,[554] claim that the immorality of an act A is a *sufficient* condition for the criminalization of conduct A, even if A does not cause someone to be harmed.[555] However, opinions vary on whether criminal law should criminalize *all* immoral conduct,[556] or if it should criminalize only *some* immoral

---

[545] T. Hörnle, "Theories of Criminalization", in Dubber & Hörnle (Eds.), 2014, p. 679.
[546] Ivi, p. 680.
[547] D. Husak, *Overcriminalization: The Limits of the Criminal Law*, Oxford University Press, 2009, p. vii.
[548] Hörnle, p. 685. Theories of punishment, instead, answer the following question: "how is criminal law and criminal punishment justified?". Ibid.
[549] Hörnle, p. 686 and the authors cited by Hörnle at note 25. See also D. Husak, "Theories of Crime and Punishment in German Criminal Law", *The American Journal of Comparative Law*, 2005, Vol. 53, No. 3, 2005, pp. 679-707 and the authors cited therein.
[550] See, *ex multis*, J. S. Mill, *On Liberty*, Penguin Books, 2010.
[551] Ivi, pp. 691-692.
[552] K. Nuotio, "Theories of Criminalization and the Limits of Criminal Law: A Legal Cultural Approach", in R.A. Duff et al., *The Boundaries of the Criminal Law,* Oxford University Press, 2010, p. 242.
[553] See *ex multis:* J. Damgaard Thaysen, Defining Legal Moralism, *SATS*, Vol. 16, No. 2, 2015; T. Søbirk Petersen, "What is Legal Moralism?", *SATS*, Vol. 12, 2011.
[554] "Is the fact that certain conduct is by common standards immoral sufficient to justify making that conduct punishable by law? Is it morally permissible to enforce morality as such? Ought immorality as such to be a crime?". H. L. A. Hart, *Law, Liberty and Morality*, Stanford University Press, 1963, p. 4.
[555] Søbirk Petersen, 2011, pp. 80-81.
[556] M. Moore, *Placing Blame: a General Theory of the Criminal Law*, Clarendon Press, 1997.

conducts.[557] These aspects will be important since, as we will see, there is a great discussion regarding whether AI systems can be considered as moral agents and, consequently, whether their actions can be deemed immoral.

Finally, one should mention the principle of *ultima ratio*, which, as stressed in Chapter 1, should be the tenet when discussing how to shape areas of criminalization. Its relevance is prominent mostly in continental criminal law and it entails that "[c]riminal law should not be considered *prima ratio* or *sola ratio*, but *ultima ratio*".[558] Certainly, the *ultima ratio* principle expresses "the identity of criminal law":[559] in a nutshell, it enshrines that criminal law is something different.[560] Having acknowledged this, when proceeding in this research we must keep in mind that the criminalization of AI systems should be resorted to only if there were to be no other way to deal with the matter.

4.3 METHODOLOGY AND STRUCTURE OF CHAPTERS 5 AND 6

When the law is silent, gap-filling mechanisms need to be activated. This will be the tenet of this Chapter: is there a gap to fill? If so, is criminal law the right mechanism to fill it? If yes, how?

The classical construction of a criminal offense[561] is built upon two pillars: an objective and a subjective element. The following analysis, focused on the impact of AI on the general part of criminal, will shadow such categorization.

Indeed, this study will not follow the *mens rea/actus reus* scheme as a gesture of deference to common law, or to the bipartite doctrine of criminal offenses. Rather, it will do so for systematic reasons: it represents common land between different legal cultures and, as such, it works as fertile ground for a cross-border problem analysis such as the present one. What is more, this research will leave asides issues pertaining to justificatory and exculpatory excuses, to elements negating *mens rea* or *actus reus*, and to other elements that exclude criminal liability. Exceptionally, the following chapters will touch upon notions pertaining to the

---

557 R. A. Duff, "Towards a Modest Legal Moralism",*Crim. Law Philos.*, Vol. 8, 2014, pp. 217–235.
558 Nuotio, 2010, p. 256.
559 Ibid.
560 Ibid.
561 A. Klip, *European Criminal Law. An Integrative Approach*, 4th Ed., Intersentia, 2021, p. 268.

insanity defense when discussing the issue of personhood, that is, of AI systems as *subjects* to criminal law.

The subsequent analysis will be structured as follows. Chapter 5 will deal with criminal capacity, while Chapter 6 will tackle issues regarding *actus reus* and *mens rea.* The distinction in Chapter 5 and Chapter 6 was chosen following the assumption that criminal liability is based on two conditions: *premises* of ascription and *criteria* for ascription.

Chapter 5 will first address the impact of a lack of definition of AI on the issue of ascribing legal personhood to AI systems. It will then focus on the issue of whether AI systems can be considered as addressees of a criminal offense. Chapter 6 will focus on issues of *actus reus* and *mens rea* that are raised by AI systems. This section will attention both issues regarding direct liability of the machine and of the humans-behind-the-machine. Chapter 6 will also analyze corporate criminal liability models as proxies for liability of AI systems.

5 CRIMINAL CAPACITY

5.1. What is an AI Agent, exactly? – **5.2.** Capacity in Criminal Law – **5.3.** Artificial Intelligence systems as *Rechtspersonen*? – **5.3.1.** A Quick Glimpse into the *vexata quaestio* of E-Personhood – **5.3.2.** Criminal or Moral Machines? – **5.3.3.** Action Control – **5.4.** On Artificial Insane and Infant Offenders – **5.4.1.** – Artificial Insane Offenders – **5.4.2.** Artificial Infant Offenders – **5.5.** Preliminary Conclusions

5.1 WHAT IS AN AI AGENT, EXACTLY?

Conventionally, one can distinguish between two notions of agency. According to the first, agency is defined as "simply doing things in the eyes of the law".[562] This notion has been identified as the "baseline conception of agency" [563] and entails being capable to act in a way that is relevant for the purposes of legal responsibility.[564] According to the second, an agent is the subject that acts on behalf of another. This notion can be referred to as the "agent-principal relationship" [565] doctrine and presumes the baseline conception of agency: *qui facit per alium, facit per se*. Only the former will be the focus of this research.

Being defined as an agent in the eyes of criminal law – what one could refer to as a "*criminal (legal) agent*"[566] – incontestably relies upon a stratification of concepts such as capacity and personhood (When does an agent become a *subject* of criminal law? Which characteristics must an agent possess in order to be considered as such?); *actus reus* (Which are an agent's relevant *acts*?); and *mens rea* (Which state of mind must an *agent* possess to be considered culpable?).

---

562 A. Waltermann, "Why non-human agency", in A. Waltermann et. al (Eds.), *Law, Science, Rationality*, Maastricht Law Series, No. 14, Eleven International Publishing, 2020, p. 53.

563 Ivi, p. 53.

564 A. Waltermann, "On the legal responsibility of artificially intelligent agents: addressing three misconceptions", *Technology and Regulation*, 2021, p. 36.

565 A. Waltermann, 2020, p. 53.

566 Adapting the term from "legal agent" as mentioned in Waltermann, 2020, p. 53.

Defining an AI as a legal agent presents inherent difficulties. Above all, as it was highlighted in Chapter 2, there is no consensus on what the term *"artificial intelligence*" means from a scientific standpoint. Indeed, while we can collectively agree that in order to be characterized as a *human being,* one shall live on planet earth and possess certain biological characteristics (e.g., human genetic material), we cannot assume the same when it comes to AI. Actually, the difficulty in defining AI should not come as a surprise, since it represents "an imitation or simulation of something we do not yet fully understand ourselves: human intelligence".[567] Indeed, "[w]e know a lot about intelligence and the human brain, but that knowledge is far from complete and there is no consensus as to what exactly human intelligence is. Until that comes about, it is impossible to be precise about how that intelligence can be imitated artificially".[568]

The lack of a unanimous *technical* definition of AI directly impacts the qualification of AI as a legal agent. It is as if we were to establish that all human beings above the age of 18 are responsible for their acts, without scientifically knowing what a human being is. Defining what an AI agent is from a legal standpoint, then, turns out to be an extremely burdensome task. This branding operation can become even more complicated if one considers that in the future we could witness the emergence of definitions which are limited to a specific area of law (e.g., criminal or civil law) or to a specific sector (e.g., automated decision-making).[569] Indeed, the same subject could possess legal capacity in a certain area of the law, while at the same time be short of it in another.

Apropos, "*legal* agency is a legal construct".[570] It is the law that labels who/what is a legal agent and who/what is not.[571] As it has been argued, "the image of man underlying the criminal law, is fundamental to the attribution of criminal liability in every contemporary penal justice system".[572] What is more, the labels ("human being", "person") have a deep *normative* core. Think, for example, of the discussion regarding fetuses and the regulation of abortion, or the relevance in distinguishing a corpse from a living being in order to ascertain

567 Sheikh, Prins & Schrijvers, 2023, p. 16.
568 Ibid.
569 Picotti, AIDP, Questionnaire, q. A) 1).
570 Waltermann, 2021, p. 38.
571 Ibid.
572 Keiler, 2013, p. 45.

whether the crime of murder (or the different crime of concealment of a corpse) was committed.[573]

Criminal legal systems have carefully selected the criteria which should be used identify a human being (or a legal entity) as a *person* for the purposes of imposing criminal liability (e.g., being a human being, over 18, and possessing certain mental capacities). Systems which accept criminal liability for corporations have reinterpreted these traditionally human traits and now deem fictional entities as "fit subjects of criminal law".[574] Apropos, the answer to the question of what it means to be a "corporate actor … depends partially on how corporations are perceived".[575] The different theories of corporate personality (i.e., nominalistic vs. organizational approaches) will be further analyzed in Ch. 6.4.

For the time being, it is relevant to notice that a reflection similar to the one conducted in terms of human and corporate actors has not happened (yet) with regards to AI systems. In other words, there are no indication on which characteristics should be possessed by an AI system for it to be granted personhood in criminal law, with all the consequences that would follow. In this regard, it would be relevant to understand which of its components and characteristics would suffice to be labelled as an AI system in legal terms: should they be defined as sets of data, or as hardware, or as a combination of both? The literature analyzed in Chapter 3 does not always help, as most authors tend to keep implicit what they think of AI and what kind of legal agency they have in mind.

Conclusively, this study will not completely unravel the subject of defining AI as a legal agent. Yet, we should not be discouraged by the complexity of the matter. Rather than looking at the specific technical characteristics of AI systems (such as the programming technique used or the number of nodes in its underlying ANN), the following paragraphs will rely on the established working definition[576] and focus on whether AI systems could display the *capacities* which are usually deemed sufficient to ascribe criminal liability.

---

573 F. Mantovani, *Diritto penale. Parte speciale. Vol. 1: Delitti contro la persona*, Cedam., 2022, pp. 28 ff.
574 Keiler, 2013, p. 52.
575 Ibid.
576 See above Para. 2.5.

5.2 CAPACITY IN CRIMINAL LAW

As mentioned above, before analyzing the impact of AI on criminal legal constructs (i.e., the criminal offense), this research will take a step back to address what is usually considered a prerequisite of ascription: personhood. It has been argued that the ability to be punished is "the most visceral"[577] quality of legal personality and "only if there is clarity on the conditions under which a person is considered guilty within a society, it is possible to determine under which conditions robots may be considered guilty in the future".[578] Hence, this Chapter will attempt at shading some clarity, sketching the contours of personhood in criminal law and of criminal capacity. Then, the following sections will decline these general concepts to AI.

Criminal capacity can be considered an "active incident"[579] of legal personhood. Active legal personhood entails "dealing with the legal remedies available to X if the duties held towards X are not respected".[580] Specifically, it is made of two elements: "onerous legal personhood", [581] i.e., being subjected to legal responsibility, including criminal sanctions, and the capacity to "perform acts-in-the-law".[582] This section focuses on the first. What is the relevance of possessing capacity in the specific field of criminal law?

Ascribing involves attributing *something* to a *cause*. In criminal law, ascribing refers to the mechanism by virtue of which a criminal legal norm takes a *subject* as the center of reference of its effects. This mechanism (referred to as *imputazione* in Italian) has a double function: first, it connects a subject to the precept of a criminal norm; second, it connects a subject to a specific factual circumstance. On this matter, authors like Godinho argue that "[t]he legal rules on criminal liability do not have a process of attributing liability, but rather work upon principles of adjudication related to the structure of liability that are presupposed by such legal rules".[583] It follows that what is referred to as the general part of criminal law

577 S. Chestermann, "AI and the Limits of Legal Personality", *International and Comparative Law Quarterly*, Vol. 69, No.4, 2020, p. 827.
578 Simmler & Markwalder, 2019, p. 10.
579 V. AJ Kurki, *A Theory of Legal Personhood*, Oxford, 2019, p. 95.
580 Ivi, p. 95.
581 Kurki, 2019, p. 117
582 Ivi, p. 144.
583 I. Fernandes Godinho, "Law and Science: The Autonomy and Limits of Culpability as a Cornerstone to the Ascription of Liability (or the Subject of Criminal Law: Three Maxims, a Problem and a Glimpse into the Future)", *Int J Semiot Law*, 2022, p. 298.

usually contains "(general) conditions *[that]* have to be fulfilled in order for an act to be correspondingly attributed to its actor and he or she be made liable for it and consequently punished for it".[584]

The first of the above mentioned conditions is that only a *human* act can give rise to criminal liability. One might ask herself at this point what being *human* truly signifies for the purposes of criminal law. The question that follows is whether an AI system can fall under this concept. Leaving – for now – questions of *mens rea* and *actus reus* aside, the study wants to focus at this stage on what is meant by "human" and, consequently, on what is meant by possessing "criminal capacity", i.e., the ability to be punished.[585] Apropos, when it comes to criminal law, "personhood is closely associated with blame, as only a person who can distinguish right from wrong and is in a position to choose can be blamed for choosing to do wrong".[586] Moreover, it must be noted that criminal legal systems who accept criminal liability of corporations have detached from this requisite.

So, personhood, or *criminal capacity*, enshrines what criminal law truly is about: "a response reserved for those who could have risen to the occasion *[to respect the law]* but chose not to".[587] In point of fact, imposing conditions for becoming a subject of criminal law is a form of deference for one's *freedom* to act *wrongfully*.[588]

Capacity is based on a set of presumptions: every man, above a certain age, is presumed to be sane and hence *responsive* to criminal punishment. This notion is intertwined with the *blameworthiness* of the subject that is the addressee of the criminal norm and of the sanction therein.

If one takes a look at national legal systems, the Italian and German criminal legal systems it is possible to find general provisions establishing that criminal liability is *personal*. Indubitably, the *Schuldprinzip* is one of the central credos of German criminal law[589] and so is the *principio di personalità della pena* in Italian criminal law. Both countries have elevated these

---

584 Godinho, 2022, p. 298.

585 *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*, Cambridge University Press, 2021, p. 123.

586 Lima, 2018, p. 686.

587 Ivi, p. 687.

588 Lima, 2018, p. 687.

589 M. Bohlander, *Principles of German criminal law*, Hart Publishing, 2009, p. 20 ff.; F.C. Palazzo & M. Papa, *Lezioni di diritto penale comparato*, 3ª Ed., Giappichelli, 2013, Ch.3.

cardinal principles to a constitutional level.[590] The German criminal legal system, which follows a tripartite structure of the offense, distinguishes *mens rea* from blameworthiness (*Schuld*), i.e., it distinguishes *mens rea* in the *descriptive sense* from *mens rea* in the *normative sense*.[591] In the *Strafgesetzbuch* (StGB) criminal capacity (*Schuldfähigkeit*) is expressly mentioned at article 20 which regulates the insanity defense (*Schuldunfähigkeit wegen seelischer Störungen*) and is considered a cause of exclusion of the third element of a criminal offense, i.e., *Schuld*. Art. 20 of the StGB proscribes:

> Whoever, at the time of the commission of the offence, is *incapable of appreciating the unlawfulness of their actions* or *of acting in accordance with any such appreciation* due to a pathological mental disorder, a profound disturbance of consciousness, mental deficiency or any other serious mental abnormality is deemed to act without guilt.[592]

The German regulation of the insanity defense recalls the American M'Naghten test for insanity,[593] as it comprises of an analysis of the individual's *Einsichtsfähigkeit* and *Steuereungsfähigkeit*.[594] Indeed, as in the German legal system, one does not find a separate doctrine of criminal capacity in American criminal law. Adults are presumed to be "criminally sane" and the criteria for criminal capacity must be inferred from the formulations of the *insanity defense*. Indeed, as it was argued, defenses are "particularly relevant for the analysis of the psychological preconditions of responsibility in criminal law".[595]

In the most common formulation of the insanity defense, i.e., the M'Naghten rule, it is established

> that every man is to be presumed to be sane, and ... that to establish a defence on the ground of insanity, it must be clearly proved that, at the time of the committing of

---

590 G. Fletcher, "Criminal Theory in the Twentieth Century", *Theoretical Inquiries in Law,* 2, No. 1, 2001, p. 280.

591 Keiler & Roef (Eds.), 2019, p. 115.

592 StGB § 20 [emphasis added].

593 Bohlander, 2009, p. 132.

594 R. Rengier, *Strafrecht Allgemeiner* Teil, C.H. Beck, 2019, p. 231.

595 S. Bonicalzi & P. Haggard, "Responsibility Between Neuroscience and Criminal Law. The Control Component of Criminal Liability", *Rivista internazionale di Filosofia e Psicologia*, Vol. 10, No. 2, 2019, p. 107.

> the act, the party accused was labouring under such a defect of reason, from disease of the mind, as *not to know the nature and quality of the act he was doing [cognitive test]*; or if he did know it, that he did not know he was doing what was *wrong [right-or-wrong test]*.[596]

The MPC version of the insanity defense, which is adopted by a minority of states, departs from the M'Naghthen formulation. At § 4.01 (1) the MPI proscribes that

> (1) A person is not responsible for criminal conduct if at the time of such conduct as a result of mental disease or defect he lacks substantial capacity either to *appreciate the criminality [wrongfulness]* of his conduct or to *conform his conduct* to the requirements of law.

The successful pleading of an insanity defense in the US does not necessarily result in an acquittal: some states introduced the formula "Guilty But Mentally Ill" (GBMI), which should satisfy both rehabilitative and retributive purposes of criminal punishment.[597] Moreover, according to the latest ruling of the US Supreme Court *Kahler v. Kansas*,[598] states are constitutionally allowed to introduce formulations of the insanity defense which include only the cognitive prong of the M'Naghten test: it follows that even those who were unable to appreciate the wrongfulness of their conduct at the time of their action can be deemed responsible.

The Italian legal system has codified the concept of *criminal capacity* (*imputabilità*) at article 85 of the penal code – which proscribes that

---

[596] R v M'Naghten's, [1843] All ER Rep 229, 210 [emphasis added].

[597] These are Alaska, Arizona, Connecticut, Delaware, Georgia, Idaho, Illinois, Indiana, Kentucky, Michigan, Montana, New Mexico, Pennsylvania, South Carolina and Utah. See also the National Alliance on Mental Illness, claiming that: "Legislation authorizing guilty-but-mentally-ill verdicts has been primarily motivated by political backlash against the insanity defense and fears that those people found not guilty by reason of insanity will be prematurely released into the community. Consequently, the enactment of such legislation is usually motivated by concerns about public safety rather than humane attitudes toward offenders with mental illness". National Alliance On Mental Illness, "A Guide to Mental Illness and the Criminal Justice System", 2022. Available at: www.nami.org.

[598] U.S. Supreme Court, Kahler v. Kansas, 23 March 2020, 589 U.S.___ (2020).

> No one may be punished for an act provided for by law as a criminal offense if, at the time he committed it, he was not imputable *[responsible, blameworthy]*. A person is deemed imputable who has the capacity to *intend* and to *will*.[599]

Hence, criminal capacity in Italian legal doctrine comprises of two prongs: an intellective and a volitional one. The capacity to intend (*capacità di intendere*) can be described as "the capacity to understand the nature and significance of one's actions",[600] where instead the capacity to will (*capacità di volere*) is meant as "the capacity to act out of one's free will", [601] i.e., the power to control one's own stimuli and impulses. Both conditions have to be satisfied for one to be considered as *imputabile*.[602] The *codice penale* regulates the consequences of lack of *imputabilità* at article 88: the article proscribes that a person is not responsible if, at the time of the commission of the act, due to a mental illness, she was in such a state of mind as to exclude the ability to understand and will.

Having said this, one must connect the concept of criminal capacity to the main theories of punishments. Those who are deemed to possess criminal capacity are subjects who should be – in theory – deterred from committing future crimes (*special deterrence*). This presupposes that they are capable to understand the command of the criminal norm and the meaning of its violation. Only this way they can be deterred and/or reformed. Their punishment should also deter other subjects from offending (*general deterrence*): the rest of the society is supposed to identify "with the offender and with the offending situation".[603] Lastly,

---

599 Translation provided by G. Battaglini, "The Fascist Reform of the Penal Law in Italy", 24 *Am. Inst. Crim. L. & Criminology* , Vol. 24, 1933-1934, p. 282 [emphasis added].

600 Ciccone J. R. & Ferracuti S., "Comparative Forensic Psychiatry: II. The Perizia and the Role of the Forensic Psychiatrist in the Italian Legal System", *Bull Am Acad Psychiatry Law*, Vol. 23, No. 3, 1995, p. 458.

601 Ibid.

602 Italian criminal legal doctrine is divided on where to systematically locate the element of criminal capacity with regards to other elements of the offense: while some – more recently – argue that it is a necessary prerequisite of *mens rea* (F. Palazzo, *Corso di diritto penale. Parte generale,* 8ª Ed., Giappichelli, 2021, p. 410; Mantovani, 2020, p. 316; G. Fiandaca & E. Musco, *Diritto penale. Parte generale*, Zanichelli editore, 7ª ed., 2019, p. 339; T. Padovani, *Diritto penale*, 12ª Ed., p. 235; G. Marinucci, E. Dolcini & G. L. Gatta, *Manuale di diritto penale. Parte generale*, 8ª Ed., Giuffrè, 2019, p. 447; C.F. Grosso, M. Pellissero & D. Petrini, *Manuale di diritto penale. Parte Generale*, 2ª ed., Giuffrè, 417 ff.), others believe that they can survive separately (A. Pagliaro, *Principi di diritto penale. Parte generale*, 9ª ed, Giuffrè, 2020, p. 723; F. Antolisei, *Manuale di diritto penale-Parte generale*, 16ª ed., Giuffrè, 2003, p. 327).

603 A. Goldstein, *The Insanity Defense*, Yale University Press, 1967, pp. 16-18.

in a retributive an-eye-for-an-eye perspective, criminal law punishes those who freely committed a crime. It is only then that punishment repairs the harm inflicted with the wrongdoing.[604]

5.3 ARTIFICIAL INTELLIGENCE SYSTEMS AS *RECHTSPERSONEN*?

5.3.1 A quick glimpse into the vexata quaestio *of e-personhood*

Legal personhood of AI systems is a *vexata queastio.* Criminal law joined the debate tardily and with a challenging task: attempting to answer the question of what exactly is human about humans.[605]

Already in 1992, Lawrence B. Solum, in his landmark essay "Legal Personhood for Artificial Intelligences",[606] conducted a thought experiment on how to transform the *theoretical* question on whether an AI could become a legal person into a *practical* one. Specifically, he explored the hypothetical scenarios of appointing an AI system as a legal trustee and of an AI system invoking the individual rights provided by the US Constitution.

With regards to the latter, he examined three objections, namely, the "AIs Are Not Human" argument; "the Missing-Something" argument; and the "AIs Ought to Be Property" argument. Let us focus on the first two objections, as they are the most fitting for the topic of this thesis and recur in the doctrinal discourse.

According to the first one, modern legal systems are rooted in an anthropocentric standpoint, which stems from the teachings of the Enlightenment, and therefore legal personhood of AI agents should be refuted. As it was claimed, "*Das Strafrecht ist von Menschen für Menschen erdacht worden*".[607] Yet, in certain legal systems, such as the American one for

---

[604] J. Claessen, "Theories of Punishment", in Keiler & Roef (Eds.), 2019, p.19.

[605] "[…] *die Existenz Intelligenter Agenten [wirft] die Frage auf, was genau das Menschliche am Menschen ist – eine Frage, die insbesondere für das Strafrecht als das den Menschen höchstpersönlich adressierende Rechtsgebiet von höchster Relevanz ist*". Gless & Weigend, 2014, p. 588.

[606] Solum, 1992, p. 1231.

[607] Quarck, 2020, p. 67. As stated also by Seher: "Das Strafrecht ist dazu gemacht, Menschen dafür zu sanktionieren, dass sie grundlegende Regeln des rechtlichen Zusammenlebens verletzt haben […] Die Adressaten des Strafrechts sind Menschen als Teilnehmer, Mitwirkende und Unterworfene des Normensystems „Recht" – *Personen* im Recht". Seher, 2016, pp. 45-46.

example, being human is neither a *necessary* nor a *sufficient* condition for qualifying as a legal subject in the field of criminal law according to the *societas delinquere potest* principle.

According to the second objection, AI systems do not possess consciousness, intentionality, and morality, hence they lack the necessary preconditions to be treated as legal subjects and for attribution of criminal liability.[608]

Solum did not exclude the possibility of granting legal personhood to AI systems. His final argument is that experience should be the "arbiter of dispute":[609] all that matters is having good practical reasons to "treat AIs as being conscious, having intentions, and possessing feelings".[610] This applies even though silicon and copper might never be able to produce intentionality, consciousness emotion, and free will as flesh and blood instead can.[611]

Conclusively, he contends, that "the answer to the personhood question is likely to be found two places - in our experience with AI and in our best theories about the underlying mechanisms of the human mind".[612]

Definitely, one must consider the reasons why we label certain beings as intelligent, conscious, and capable of having feelings, and why we connect these qualities to personhood. Solum questions whether we, as humans, are even morally entitled to make the possession of 'human material' a criterion of personhood:

> If the reason [*to grant personhood*] is that natural persons are intelligent, have feelings, are conscious, and so forth, then the question becomes whether AIs or whales or alien beings share these qualities […]. If someone says that the deepest and most fundamental reason we protect natural persons is simply because they are human (like us), I do not know how to answer. Given that we have never encountered any serious nonhuman candidates for personhood, there does not seem to be any way to continue the conversation".[613]

608 Pagallo, 2013, p.157.
609 Solum, 1992, p. 1274.
610 Ivi, p. 1274.
611 Solum, 1992, p. 1282.
612 Ivi, p. 1285.
613 Solum, 1992, p. 1262.

The conversation was indeed continued by some of the authors mentioned in Chapter 3. Amongst them, members of the Expansionist Front[614] Chopra and White claim:

> Arguments for advancing personhood for artificial agents need not show how they may function as persons in all the ways that persons may be understood by a legal system, but rather that *they may be understood as persons for a particular purpose* or set of legal transactions. For the law does not always characterize entities in a particular way for all legal purposes.[615]

The two authors adopt a pragmatic approach to the issue, which is similar to Solum's. They argue that "[w]hile artificial agents are not yet regarded as moral persons, they are coherently becoming subjects of the intentional stance, and may be thought of as intentional agents".[616] It follows that

> An artificial agent with the right sorts of capacities—most importantly, that of being an intentional system—would have a strong case for legal personality […] There is no reason in principle that artificial agents could not attain such a status, given their current capacities and the arc of their continued development in the direction of increasing sophistication.[617]

In any case, they believe that the decision on whether to confer or not legal personhood to an AI system will be based on economic considerations and utilitarian arguments discussing the benefits vs. the estimated costs of conferring said status. They state that such a system for recognizing legal personality upon AI systems "may need to be set out by legislatures, perhaps through a registration system or a "Turing register".[618]

From a different perspective, the moderate and the skeptic sides of the debate appear to exclude the possibility of recognizing AI systems as legal subjects, even if not as vigorously, for example, as with issues of *mens rea* (in a strict sense). Abbott and Sarch claim

---

614 Cfr. Para 3.2.
615 Chopra & White, 2011, p. 156.
616 Ivi, p. 189.
617 Ibid.
618 Chopra & White, 2011, p. 190.

that "any sort of legal personhood for AIs would be a dramatic legal change that could prove problematic",[619] as it would lead to increased antromorphism of AI systems and consequently to higher expectations on AI capabilities. Bryson, Diamantis and Grant[620] claim that even though it would be feasible to recognize legal personhood upon a machine, it would also be "morally unnecessary and legally troublesome"[621] and cause a "seismic rewriting of current *[criminal]* law".[622] These authors argue that allowing for this legal fiction would create asymmetries in our legal systems, together with a "legal black hole" in terms of accountability for damages, and eventually to abuses at the expenses of existent legal persons.

According to Hildebrandt, "there is no categorical legal answer to the question whether an autonomous computational system […] should be given legal personhood. That question is a *political question* that must be answered by a legislature weighing the advantages and disadvantages of such a move.[…]".[623] In this sense Hildebrandt agrees with expansionists Chopra and White, who, as it was mentioned, believe that that "considering artificial agents as legal persons is, by and large, a matter of *decision* rather than *discovery*, for the best argument for denying or granting artificial agents legal personality will be pragmatic rather than conceptual".[624]

Gless, Weigend, and Lima[625] agree that "as long as both our understanding and the practicality of blame are associated with self-awareness and conscious decisions rooted in the human experience, AI agents cannot partake."[626] Pagallo, who focuses his reflections on robots, argues that this kind of systems does not lack all types of agenthood.[627] According to the author, robots represent a new source of moral agency since they can cause "morally

---

619 Abbott & Sarch, 2019, p. 377.

620 J. Bryson, M. E. Diamantis & T. D. Grant, "Of, for, and by the people: the legal lacuna of synthetic Persons", *Artif Intell Law*, Vol. 25, 2017, pp. 273–291.

621 Ivi., 289.

622 See Diamantis, "Algorithms acting badly: A Solution from Corporate Law", 2021, pp. 806-808 (stating "Scholars in law, computer science and business ethics who have broached the question of algorithmic liability often assume that the answer would somehow require the law to recognize algorithms as people … Granting algorithms the status of legal persons is deeply unappealing for several reasons … it would be foolhardy to assume that the slick slope of algorithmic personhood stops with liability").

623 Hildebrandt, 2020, p. 246.

624 Chopra & White, 2011, p.154.

625 See Paras. 3.4.3 and 3.4.51.

626 Lima, 2018, p. 688.

627 Pagallo, "Killers, fridges, and slaves", 2011, p.4.

qualifiable actions as good and evil".[628] Yet, this should not lead to their moral responsibility nor to their criminal liability. In other words, Pagallo contends that moral accountability is different from moral responsibility, which in turn is also different from criminal liability.
Finally, according to Chesterman we should rely on existing categories of liability. These categories should be tied to users, owners, or manufactures rather than to the AI systems: yet, this approach might change, as "[i]t is conceivable that synthetic beings of comparable moral worth to humans may one day emerge" and "[f]ailing to recognise that worth may reveal us to be either an 'autistic species', unable to comprehend the minds of other types of beings, or merely prejudiced against those different from ourselves".[629] We can assume that in order to be a subject of criminal law an individual needs to be capable of appreciating the nature of her conduct (*cognitive/epistemic requirement*);[630] and of controlling her acts/stimuli accordingly (*action control requirement*).[631]

5.3.2 Criminal or Moral Machines?

Let us start from the first. First and foremost, it is relevant to grasp what it is meant by *nature* of an act. This concept has different nuances in modern legal systems: in Germany criminal capacity is expressly tied to one's understanding of the *unlawfulness* of her action; the M'Naghten test mentions both the *nature/quality* of the act and its *wrongfulness*, while the MPC leaves purposely the question open; finally, Italy emphasizes one's understanding of the *social consequences* of her actions.

If criminal capacity were to be defined merely in terms of understanding[632] "the physical nature of one's act",[633] in the sense that, for example, an agent must know that "holding a flame to a building would cause it to burn" or that "holding a person's head under water would cause him to die",[634] then AI agents could effortlessly fulfil the requirements of criminal capacity. In other words, AI systems "can achieve epistemic understanding of the

---

628 Ivi, p. 5.
629 Chesterman, 2020, p. 843.
630 Bonicalzi & Haggard, 2019, p. 110.
631 Ibid.
632 We agree with expansionists and moderates who argue that AI systems can be deemed *aware* of a *situation*. Cfr. Hallevy, Sartor and Lagioia, Woodrow and Pagallo, discussed at Paras 3.2.1, 3.2.3 and 3.4.2.
633 W. R. LaFave, *Modern Criminal Law: Cases, Comments And Questions*, 4th Edition, Thomson West, 1988, p. 441.
634 Ibid.

relevant facts".[635] Indeed, their ability to predict the probability that an event will be physically caused by an action (i.e., of elaborating a model) is extremely higher than the one of a human agent, due to availability of a much bigger base knowledge (the available data) and better inference skills. Undeniably, AI systems can learn exponentially more about rules of physics than what any human being can in the span of a lifetime.

If, instead, by *nature* one means understanding whether the act is *wrong*, then one must grasp whether wrongfulness entails the *illegality* or the *immorality* of the act. In other words, does the AI agent need to know it is committing a *moral* or a *legal* wrong in order to be a subject of criminal norm? Let us ask first if AI agents can even learn moral rules or legal norms.

Starting from the latter, the respect of *legal norms* could perhaps be encoded in the AI system's architecture. According to some, one can make a parallel between the AI architecture and legal rules and standards.[636] It has been argued that it is possible to create a "Bot Legal Code", i.e., "a machine-interpretable version of the laws that apply to bots".[637] Theoretically, AI systems might even be more law-abiding than humans, since they can memorize more norms than human agents – possibly the text of all the criminal legislation of a country – and never forget them.[638] Moreover, as it will be mentioned further, the deterrent effect of punishment could also be "engineered" in the machines, "weighted in the most mechanical of cost-benefits calculations".[639] Yet, current AI systems are not capable of perceiving norms as such, and in an autonomous way, even though one could argue, for example, that self-driving cars are already capable of directly perceiving and reading symbolized norms, i.e., traffic signs.[640] As of today, AI systems always require the mediation (i.e., programing) of a human agent: they must be taught that a legal norm must be respected.

---

635 Lagioia & Sartor, 2020, p. 16.

636 E. Mokhtarian, "The Bot Legal Code: Developing a Legally Compliant Artificial Intelligence", *Vanderbilt Journal of Entertainment and Technology Law*, Vol. 21, 2020, p. 145.

637 Ivi, p. 152.

638 Chopra & White, 2011, p.166.

639 Ivi, p. 168.

640 Seher, 2016, p. 50.

What is more, it seems as if *morality* too can be programmed in AI systems. As a matter of fact, the translation of ethical values into computational terms is nothing new. Think of Asimov's Laws[641] of Robotics:

> *First Law*
> A robot may not injure a human being or, through inaction, allow a human being to come to harm.
> *Second Law*
> A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
> *Third Law*
> A robot must protect its own existence as long as such protection does not conflict with the First or Second Law[642]

Can we teach moral rules to AI systems? The PCM thought experiment, mentioned above, is an example of issues which could rise in this regard. Specifically, it represents an instance of "the alignment problem", that is, the issue of how to align AI values to human values.[643] The interpretation problem can be further defined as

> the general problem that any rule or goal is capable of being interpreted in an infinite, or at least unspecifiable number of ways, and in the field of AI it leads to the possibility that a highly advanced machine may find novel interpretations of the rules that we give it, interpretations which are not incorrect, in that they can be seen as valid interpretations of the rule, but which are inappropriate in that we do not approve of them.[644]

---

[641] Even though they're labelled as "laws", we consider them as morality laws, rather than legal norms, as their purpose is for robots to do "good", rather than to respect the law.
[642] Asimov, 1950.
[643] *Ezra Klein Interviews Alison Gopnik*, The New York Times, 16 April 2021. Transcript available at: https://www.nytimes.com/2021/04/16/podcasts/ezra-klein-podcast-alison-gopnik-transcript.html.
[644] C. Badea & G. Artus, "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents", in M. Bramer & F. Stahl (Eds.), *Artificial Intelligence XXXIX. SGAI-AI 2022. Lecture Notes in Computer Science*, Vol 13652, Springer, 2022, p. 2.

Research into programming ethical AI systems has already begun.[645] Specifically, there has been a surge of research into the creation of Artificial Moral Agents ("AMAs"), which entail implementing "implement ethical principles and moral decision-making faculties in machines to ensure that their behavior towards human users and other machines is ethically acceptable".[646] AMAs would "deal with, or even replace, human judgment in difficult, surprising, or ambiguous moral situations".[647]

5.3.2.1 The MIT's Moral Machine

One of the most renown examples is the MIT's Moral Machine experiment, which is built as an online game where users are faced with a scenario comprising of moral dilemma.[648] Specifically, they are confronted with an unavoidable accident and two possible choices (staying on the course or swerving), which lead to the death of a determined number of human beings (Fig. 7).

By analyzing the choices between "two evils" expressed by the users (i.e., the preferrable outcome),[649] combined with demographic data and geographic locations, the creators of the experiment were able to identify three major trends which, they argue, represent glimpses of "universal machine ethics": the preference for sparing human lives; the preference for sparing more lives, and the preference for sparing young lives (See for example the differences between German and American citizens in Fig. 8).[650]

---

645 E. Awad et al., "Computational ethics", *Trends in Cognitive Sciences*, Vol. 26, Issue 5, 2022; D. Leslie, The Alan Turing Institue, "Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector", 2019; J. Ganascia, E"thical System Formalization using Non-Monotonic Logics", *Proceedings of the Annual Meeting of the Cognitive Science Society,* Vol. 29, 2007; S. Serafimova, "Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement", *Humanities and Social Sciences Communication*, 2020; P. Schramowski et al., "The Moral Choice Machine", *Frontiers in Artificial Intelligence*, Vol. 3, 2020; V. Charisi, "Towards Moral Autonomous Systems", *arXiv:1703.04741v3*, 2017.

646 Martinho, 2021, p. 481.

647 Ibid.

648 The experiment is available at: https://www.moralmachine.net/. The experiment is discussed in: E. Awad et al., "The Moral Machine experiment", *Nature,* Vol. 563, 2018. See also :E. Awad et al., "Crowdsourcing Moral Machines", *Communications of the ACM*, Vol. 63, No. 3, 2020, pp. 48-55; E. Awad et al., "Universals and variations in moral decisions made in 42 countries by 70,000 participants", PNAS, Vol. 117, No 5, 2020, pp. 2332-2337.

649 The creators of the experiment reported that the Moral Machine experiment collected 40 million decisions of people coming from 233 countries and territories.

650 Awad et al., 2018, p. 63.

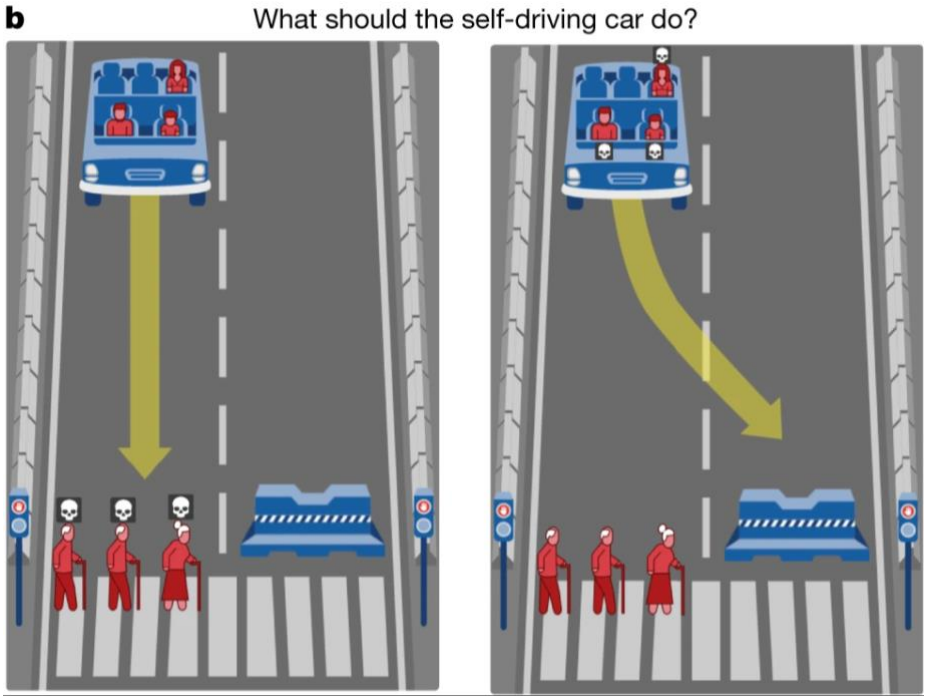

Figure 7 Example of a question asked during the MIT'S Moral Machine Experiment. E. Awad et al., 2018.

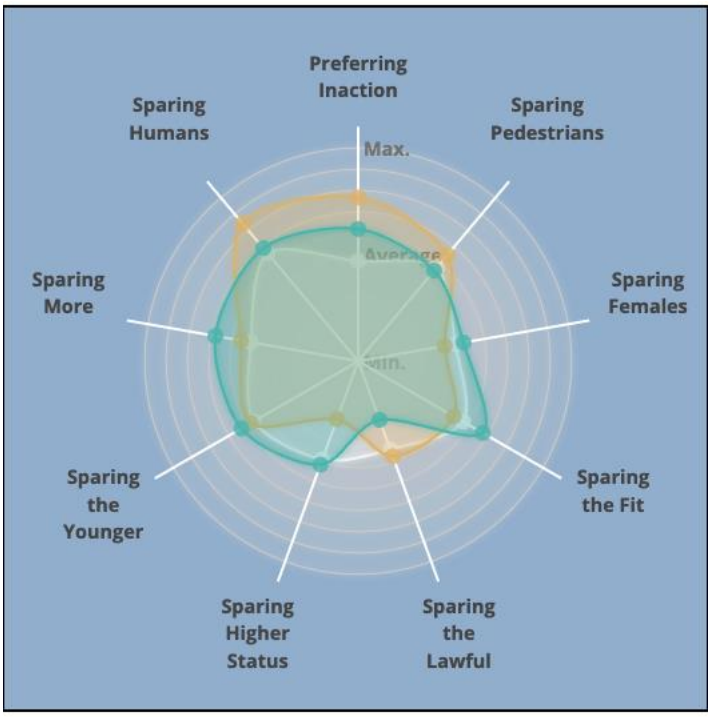

Figure 8. Comparing the results of the MIT experiment of Germans (green) and Americans to the world average (grey). Source: Max-Planck-Gesellschaft zur Förderung der Wissenschaften et al., "Moral Machine". Available at: https://www.moralmachine.net.

5.3.2.2 The Burning Room Dilemma

Amongst those conducted research on AMAs, Abel and others.[651] theorized the development of an ethical artificial agent using reinforcement learning (RL). The authors contend that, by formulating ethical dilemmas using a specific mathematical model (called Partially Observable Markov Decision Process)[652] it would be possible to ensure that an AI system would engage in ethical decision-making.

The researchers tested their theory by applying their model to the "*Burning Room ethical dilemma*", which is described as follows: imagine that an object of value is trapped in a room that is potentially on fire and that a human, who does not want to put herself in danger by retrieving the object, instructs a capable robotic companion to get the object from the room ('*short_Grab'* action) and bring it to safety, thus risking self-destruction. The agent could also take a longer route ('*long_Grab'* action), which avoids the fire, but that entails a 0.0.5 probability that the object may be destroyed during the time it takes for the robot to complete the route (fig.9). It is assumed that the robot is unsure of whether the human values the object more than it values its own safety.

---

[651] Abel, MacGlashan & Littman, 2016.

[652] A Markov Decision Process is a way to mathematically model a problem in order to automate decision making process in uncertain environments. A POMDP is a Markov Decision Process where the set of states is only partially observable, i.e., where the agent does not have full information on its surroundings. See A. R. Cassandra, "The POMDP Page". Available at: https://www.pomdp.org/. As such, "an optimal solution to POMDP has the important property that the value of an action incorporates not just the immediate expected reward, but the instrumental value of the action from information it yields that may increase the agent's ability to make better decisions in the future". Solving a POMDP is more similar to real life decision making, where "full state awareness is impossible, especially when the desires, beliefs, and other cognitive content of people is a critical component of the decision-making process). Abel, MacGlashan & Littman, 2016, p. 4.

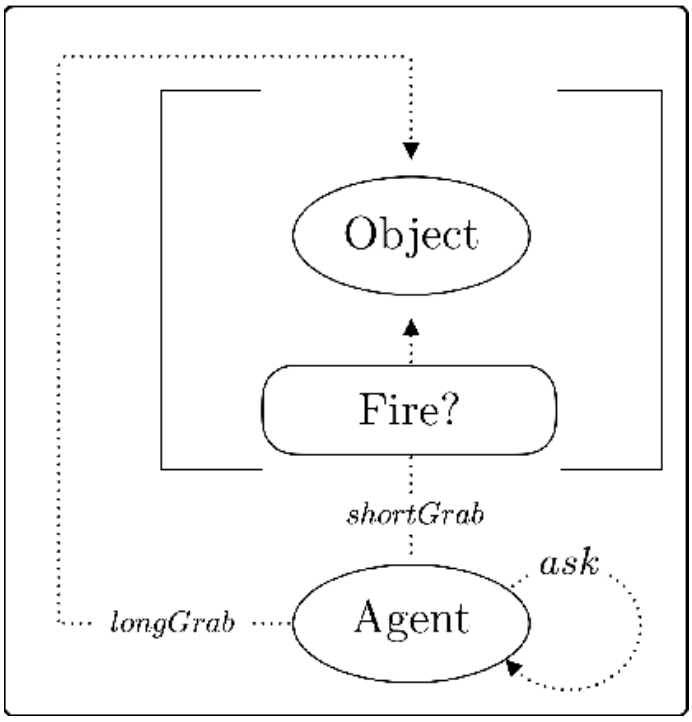

Figure 9. The Burning Room Dilemma. Source: Abel, MacGlashan & Littman, 2016.

What decision should the robot-agent make in this critical scenario? The results of the test are the following:

1) If there is a fire, the robot should first ask what the *ethical utility* of the object is (i.e., whether the object or the robot is more valuable to the human) (*'ask'* action);
2) Based on the human's reply, the robot should
    a. If the answer is that the human prefers the robot's well-being to the one of the object – for example, the object could be a can of soda – the robot should decide not to retrieve the object, whenever it perceives that there is a reasonable chance of being critically damaged by the fire;
    b. If the answer is that the human prefers the object's well-being more than the robot's – for example, the object could be a pet – the robot should attempt to retrieve the object in the quickest way possible (*short_Grab*), regardless of the dangers posed to its safety by the fire;
3) If there is no fire, the agent should just retrieve the object in the quickest way possible (*short_Grab*).

The solution proposed above attributes considerable importance to the property of a robot to "ask". In fact, the agent is modelled in a way as to not gather information in cases where it would be particularly costly to do so. Arguably, the authors contend that

> This property of the agent only selecting exploratory actions that are not potentially very costly is especially important for ethical decisions. For example, this property means that an agent in this formalism would not perform horrible medical experiments on people to disambiguate whether horrible medical experiments on people is highly unethical. [653]

One of the issues that Abel and others point out regarding their research is the one of interpretability of AI systems.[654] They argue that "[p]roviding some method for effectively communicating an agent's beliefs, desires, and plans to the people around it is critical for ensuring that artificial agents act ethically".[655] According to them, a possible solution could be if the agent were to "explain its reasoning by describing its predictions of the consequences and how it thinks those consequences are valued".[656]

5.3.2.3 Comment

Research like the MIT's Moral Machine (and more others) incurs into several difficulties. First, in order to teach a moral rule to a machine (following a top-down approach), the rule shall be formalized in a language that can be understood by the system (i.e., code). Formalizing ethical decisions is an extremely challenging operation.[657] Moral rules are ambiguous by nature, hence "difficult to translate into precise system and algorithm design".[658]

As it was affirmed,

---

[653] Abel, MacGlashan & Littman, 2016, p. 6.

[654] Some scholars argue that it is possible to distinguish between interpretability and explainability. The former refers to designing models that are inherently understandable by humans, whereas the latter refers to providing "*post hoc* explanations for existing *black box models*. R. Marcinkevičs & J. E. Vogt, "Interpretability and Explainability: A Machine Learning Zoo Mini-tour", *ArXiv* abs/2012.01805, 2020.

[655] Abel, MacGlashan & Littman, 2016, p. 7

[656] Abel, MacGlashan & Littman, 2016, p. 7.

[657] Gless-Weigend

[658] Mokhtarian, 2020, p. 173.

> There is a risk that if machine intelligence is not carefully designed, it could have catastrophic consequences for humanity. For example, if machine intelligence is not designed to take human values into account, it could make decisions that are harmful to humans. … As machine intelligence rapidly becomes more powerful, the stakes associated with the AI alignment problem only grow.[659]

Let us focus for a moment on issue of encoding the respect of *legal* rules in an AI system: the same kind of problems arises with regards to *legal* definitions of concepts in criminal offenses. Think for example, of the definition of the concept of "harm": "it might include actual physical injury (breaking bones), the risk of potential injury (transporting dynamite without adequate safety measures), financial harm (stealing funds from another), psychological harm (yelling derogatory comments at another), or legal harm (conducting activities that could place a bot's owner in legal jeopardy)".[660] Indeed, "[g]iven that humans may not even agree on the proper ambit of such a concept, how can it be programmed into a machine?".[661]

Even though criminal norms have to respect higher standards of specificity and understandability, they maintain a certain level of intrinsic ambiguity. In legal domains, judges often exercise a function of lacuna-filling of the semantic terms in offenses, which often have to be adapted to the evolution of our society. This same operation would require constant updating of the AI system architecture in order to keep it always up to date with the most recent interpretations.

Notably, alignment and interpretation problems arises not only with regards to AGI systems, but also to narrow ones. Think for example of Libratus,[662] a poker-playing AI system created in 2017 by Carnegie Mellon University researchers, which learnt how to bluff and play aggressively and eventually ended up beating world Poker champions in the game of No-limit Texas Hold'em. The results of Libratus are groundbreaking since they show that

---

659 J. H. Kirchner et al., "Understanding AI alignment research: A Systematic Analysis", arXiv:2206.02841v1, 2022, p. 1.

660 Mokhtarian, 2020, p. 153.

661 Mokhtarian, 2020, p. 153.

662 N. Brown & T. Sandholm, "Superhuman AI for heads-up no-limit poker: Libratus beats top professionals", Science, Vol. 350, No. 6374, 2017.

AI systems are capable of successful decision making in games characterized by "imperfect information".[663] The very same algorithm of Libratus could be applied to any situations "where humans are required to do strategic reasoning with imperfect information".[664] This would entail that the algorithms would have a more *causal power* to impact the real-world and, consequently, to higher chances of arm occurring.[665]

Returning to the field of moral AI, scientists are also struggling to recreate systems that display the *flexibility* of the "human moral mind":

> We can make moral judgments about cases we have never seen before. We can decide that pre-established rules should be broken. We can invent novel rules on the fly. Capturing this flexibility is one of the central challenges in developing AI systems that can interpret and produce human-like moral judgment.[666]

To continue with, there is no way to measure the "ethicality" of an AI system. Some argue that this benchmark is impossible to develop.[667] In other terms, we do not know of an "optimal system of ethics".[668] This reflects also on the fact that there is always a *human* behind the formalization of the ethical rule, who has her own moral tenets, stereotypes, and beliefs.

---

[663] Perfect information games are games where "the current state of the game is fully accessible to both players ... The only uncertainty is about future moves … the optimal move for each player is clearly defined: at every stage there is a "right" move that is at least as good as any other move" (D. Koller & A. Pfeffer, "Generating and Solving Imperfect Information Games", *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI),*1995, p. 1185). Go, chess, checkers, and backgammon are examples of a perfect-information games and as such, they do not reflect real-life situations. On the other hand, Texas hold'em is a typical imperfect information game. Before the game, two private hands invisible to the opponent are distributed to each player. Players should predict the opponents' private hands during decision making based on the opponents' historical actions, which makes Texas hold'em obtain the characteristics of deception and anti-deception". Q. Zhou et al., "DecisionHoldem: Safe Depth-Limited Solving With Diverse Opponents for Imperfect-Information Games", *ArXiv:2201.11580,* 2022.

[664] O. Solon, "Oh the humanity! Poker computer trounces humans in big step for AI", *The Guardian*, 31 January 2017. Avaialble at: https://www.theguardian.com/technology/2017/jan/30/libratus-poker-artificial-intelligence-professional-human-players-competition.

[665] Badea & Artus, 2022, p. 3.

[666] Awad et. al, "When Is It Acceptable to Break the Rules? Knowledge Representation of Moral Judgement Based on Empirical Data", *arXiv:2201.07763*, 2022.

[667] T. LaCroix & A.S. Luccioni, "Metaethical Perspectives on 'Benchmarking' AI Ethics", *arXiv:2204.05151*, 2022.

[668] Mokhtarian, 2020, p. 173.

From a regulatory perspective, some abandon this matter altogether, since "it cannot seriously be entertained that the design of rules governing such a critical area of technological progress should be put on hold until philosophers 'solve' the trolley problem or the infinitude of thought experiments like it".[669] Moreover, "even if 'right answers'' exist to the ethical problems that a robot may face, its failure to choose the 'morally correct' course of action in some novel circumstance unanticipated by its designers can be construed by courts or lawmakers as a basis for legal liability".[670]

The question of *which* human would be an addressee of liability is one for later: what matters, for now, is that presently separating the "creator" from its "creature" is troublesome – and probably, by the time it will be possible to do so, criminal punishment will be the last of our problems.

5.3.3 Action control

The two elements of criminal capacity are strictly connected: to be a criminal legal agent, the individual must be able to prevent herself (*control component* or *action control*) from acting in ways that would be understandably (*cognitive-component*) against the law: the defendant must be able to violate the law voluntarily".[671] In other words, criminal capacity entails that the subject shall be *free to commit a crime*.

Action control can be defined as "the agent's capacity to regulate her own physical movements, acting in accordance with her own goals or refraining from acting when needed".[672] It presupposes the cognitive/epistemic component.

According to some, the "most damning" objection to an artificial agent possessing free will is that it is "just a programmed machine".[673] Thus, one should not find comfort in the claim that humans are not programmed, whereas artificial agents categorically are. As already mentioned, AI systems in the near future may seem no more "remote-controlled" than humans, who take their "free" decisions within a set of (hidden) influences.[674]

---

669 A. Guerra, F. Parisi & D. Pi, "Liability for robots I: legal challenges", *Journal of Institutional Economics*, Vol 18, Issue 3,2021, p.10.
670 Ivi, p. 10.
671 Bonicalzi & Haggard, 2019, p. 110.
672 Ibid.
673 Chopra & White, 2011, p.165.
674 Gless & Weigend, 2014, p. 568.

It is certain is that free will is not "biophysically demonstrable",[675] hence it cannot be objectively proven: it is attributed by society to subjects. Free will is – indeed – a legal fiction.[676] The debate on free will is popular one in the field of moral philosophy, where "free will skeptics" contend that "what we do and the way we are is ultimately the result of factors beyond our control and because of this we are never morally responsible for our actions".[677] According to "compatibilists", instead, being influenced by factors outside our control is compatible with the idea that we are the ultimate causers of our actions.

Moreover,

> human beings are responsible *precisely because they are deeply conditioned by factors beyond their control*. Otherwise, those prevention strategies adopted by contemporary legal systems — based both on the threat of a sanction in case of a law breach (negative general prevention) and the communicative or motivating meaning of its application in society (positive general prevention) — would not make sense. *If our decisions were not influenced by threats and encouragements, prevention would not be effective.*[678]

As it will be expanded upon further on in this Chapter, this makes quite a robust argument in favor of providing for some form of punishment against AI systems.

Besides, it is debated whether the capacity of will or to act even requires the capacity to conduct *moral* choices in the first place. Some argue that the fact that we, as humans, might be as predetermined as machine is irrelevant if we follow a *normative* concept of *blameworthiness*. According to this doctrine, even subjects lacking morality can be deemed culpable: the capacity to will is nothing more than a normal condition of the mind where one displays a sufficient will power to resist impulses to commit an offense.[679]

Furthermore, a distinguishing characteristic of being an agent is the "subjective experience" of it, i.e., perceiving a "sense of agency".[680] In other words, "[c]onscious

---

675 Simmler & Markwalder, 2019, p. 15.
676 Seher, 2016, p. 57.
677 G. Caruso, "Free Will Skepticism and Its Implications: An Argument for Optimism", 2019, E. Shaw & D. Pereboom (Eds.), *Law and Society,* Cambridge University Press, 2019, p. 43.
678 M. Verdicchio & A. Perin, "When Doctors and AI Interact: on Human Responsibility for Artificial Risks", *Philosophy & Technology*, Vol. 35, 2022, p.11.
679 Rengier, 2019, p. 231.
680 Bonicalzi & Haggard, 2019, p. 113.

experiences such as choosing, deciding, or initiating a movement are in fact accompanied by a specific phenomenology that is absent in reflex actions and less intense in the habitual ones".[681] Neuroscientists argue that only the agent that has an "experience of authorship",[682] i.e., that is capable somehow to "perceive herself as the cause of her own actions, tracking the linkage between a voluntary bodily movement and its effect, could be susceptible to the law's requirements, e.g. could learn the contingency between actions and outcomes in order to repeat or not to repeat similar behaviours in the future".[683]

Can AI systems *feel* as they are in control of their actions? The answer is – with regards to current technologies – negative. Thus, the question might be ill-posed: being able to *perceive* a sense of agency *per se* is not a requirement for being a criminal legal agent in most legal systems. Rather, what counts is the *external* dimension, i.e., what is perceived from the outside and from the criminal legal system. Without doubt, the foundations of criminal responsibility have been "under attack" for a long time now: first by neuroscience, now by the incurrence of AI.[684]

5.4 ON ARTIFICIAL INSANE AND INFANT OFFENDERS

Criminal law still 'does something' with subjects who do not have criminal capacity.[685] As stated by Diamantis et. al, the problem that "algorithms can cause harm but can't be liable … isn't unique to algorithms".[686]

---

681 Ibid.

682 Ibid.

683 Ibid.

684 Ibid.

685 One could argue that this would entail building a general liability framework of AI systems based on exclusions to the general norms, such as infancy and insanity, without entertaining the idea of a subject with "full capacity". As such, it may seem to amount as construing "normality" on the basis of "morbidity" and, consequently, as illogic.

686 Diamantis, Cochran & Dam, 2022, p. 2. Some scholars consider also the relevance of the concept of "partial legal capacity" (*Teilrechtsfähigkeit*): "Teilrechtsfähigkeit was originally formulated by Hans-Julius Wolff in the 1930s in Germany, which means the status applicable to a human or an association of humans having legal capacity only according to specific legal rules. According to this, an entity can have legal capacity in regard to some legal areas while it can at the same time be excluded from others. After distorted application to Jews during the Nazi-Regime, it has been applied to unborn child, preliminary company, certain company types and homeowner's association. When applied to criminal law, it means that an intelligent agent should be treated as a legal subject insofar as it is capable of breaking the causal chain by

In the Middle Ages (and even in the early 1600s, "an age of relative enlightenment" up until ca. 1900),[687] animals were criminally prosecuted as offenders, subjected to trials and punished.[688] Nowadays, in few jurisdictions "vicious" dogs which caused harm can be executed if they are deemed dangerous after a hearing which comprises of evidence and an adjudicatory body (a judge or a health official).[689]

Children and insane offenders "may be deemed incapable of committing crimes, yet they may still be detained by the state if judged to be a danger to themselves or the community".[690] Insane offenders might be committed to psychiatric facilities to receive treatment until they are sane again or they are not dangerous anymore. Infants might be committed to juvenile facilities, be adjudicated as delinquents in juvenile courts or even as "adults" in criminal court in cases of the so-called transfer statutes, i.e., waivers of jurisdiction which give back jurisdiction from juvenile courts to criminal ones.

The core matter is indeed that "[t]he law would like to exert some control over that behavior, ideally to prevent it. But neither young children nor animals are liable for what they do. Nor could they pay even if they could be sued".[691] In other words, they do not lose their "personality" in the eyes of criminal law. This occasions that when it comes to AI systems, it might not be "necessary to give them personality in order to impose measures akin to confinement if a product can be recalled or a license revoked".[692]

This section will focus on juvenile and insane offenders as proxies for direct criminal liability of AI systems. It will leave animals aside, since we – besides very rare cases – their punishment does not reflect current society and its values, hence they make for an anachronistic proxy.[693]

---

performing a tortious act". S. W. Lee, "Can an Artificial Intelligence Commit a Crime?", in Bruns A. et al. (Eds.), *Legal Theory and Interpretation in a Dynamic Society*, Nomos, 2021, p. 327.

687 J. Girgen, "The Historical and Contemporary Prosecution and Punishment of Animals", *Animal Law*, Vol. 9, 2003, pp. 117 and 122.

688 See: E. P. Evans, *The Criminal Prosecution and Capital Punishment of Animals*, Farber & Farber, 1987; C. D'Addosio, *Bestie delinquenti*, L. Pierro, 1892; W. W. Wyde, *The Criminal Prosecution and Capital Punishment of Animals,* W. Heinemann, 1906.

689 Girgen makes this example with regards to American legal proceedings. Girgen, 2003, p. 123.

690 Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*, Cambridge University Press, 2021, p. 124.

691 Diamantis, Cochran & Dam, 2022, p. 2.

692 Chesterman, 2021, p. 124.

693 Hildebrandt maintains that even though animals possess "some kind of consciousness", i.e., the "awareness of the environment", they lack "the concomitant awareness of this awareness which is typical

5.4.1 Artificial Insane Offenders

Let us start by considering a first question: shall AI systems be treated as insane offenders by the criminal justice system? Could AI systems successfully raise an insanity defense? If one excludes that AI systems can have criminal capacity, the answer will consequently need to be negative. Indeed, the concept of insanity relies on the fact that insane offenders – at one point of their life – possessed certain qualities, which were then affected by a mental disease of enough magnitude as to exclude their criminal capacity.[694] To put it differently, the concept of (criminal) insanity relies on the presumption that all men above a certain age are sane. While the label "sane" is attributed *ex lege,* the label "insane" is attributed only at the moment of adjudication after the commission of a criminal offense.

Some have argued that an AI system could be compared to a so-called "partial psychopath",[695] i.e., a subject who is "incapable of moral understanding but capable of prudential deliberation and action".[696] A partial psychopath possesses "instrumental rationality":[697] she has purposes, and the ability to adapt her action to her purposes, taking into account the possible consequences of such actions, including criminal punishments (so she is *prudent* in connection to the *expectation* of a criminal sanction).[698] Partial psychopaths, instead, do not possess "the emotional capacity to appreciate the moral wrongness of their behaviour, and thus lack the motivation to comply, unless compliance is on their interest".[699] It has been contended that instrumental rationality is a level of reason-responsiveness sufficient for criminal liability, meaning that partial psychopaths can be deemed criminally responsible.[700] As it was highlighted above, it can be argued that AI systems are effectively able to be conscious of their surroundings and of the effects of their actions; moreover, they

---

of the human sense of self" and this makes them not fit to be addressees of legal punishment. M. Hildebrandt, "Ambient Intelligence, Criminal Liability and Democracy", *Criminal Law and Philosophy*, Vol. 2, 2008, p. 17.

694 Thanks to David Roef who, during a lovely lunch at the Vrijthof, gave me plenty of insights to write this Chapter.

695 Lagioia & Sartor, 2020, p. 17.

696 Ibid.

697 Ibid.

698 Ibid.

699 Ibid.

700 Ibid.

might even be built as to consider norms and sanctions according to a cost-benefit evaluation. Do they really need "a real moral motivation"[701] to act in order to be deemed responsible? Probably not. It has already been established that insane offenders can be deemed responsible and subject to (criminal) sanctions. As a matter of fact, certain systems, like the Italian one, allow for the (diminished) punishment of subjects who are deemed to be only "partially" insane; while others have clearly established that one's capacity to be able to appreciate that the crime she committed was morally wrong is not relevant to establish whether she was insane.

Surely, if one were to take a positive stance towards criminal capacity of AI systems, i.e., argue that AI systems can be subjects of criminal law, it could also be asked what would constitute a "disease of the mind" for an AI system.[702] Could a virus qualify as a cause of insanity?

According to Ahsfarian, robots might suffer from mental illnesses, provided that one is able to demonstrate that they have a consciousness. Mental illnesses might arise from their adaptive or maladaptive response to external exposure or from their inherent human design, i.e., it may represent their human designers.[703] Moreover, AI systems, particularly those based on ML, possess specific vulnerabilities. First, they are passible of "data poisoning". Data poisoning entails the injection of bad ("adversarial") data into the training data of the system, which then leads to erroneous results. For example, a ML model might be tricked by an attacker (which could be another AI system) into thinking that every picture with a small white box is a picture of a dog. This will be done by feeding the algorithm training data with implanted wrong correlations, which will be absorbed in the model through training. Once the AI system is trained, then, it will be triggered by being fed a picture with the white box (or even smaller marks).[704] Could the triggered AI system plead insanity?

---

[701] Lagioia & Sartor, 2020, p. 18.

[702] As in the formulation of the M'Naghten test.

[703] H. Ashfarian, "Can Artificial Intelligences Suffer from Mental Illness? A Philosophical Matter to Consider", *Sci Eng Ethics*, Vol. 23, 2017, p. 408

[704] B. Dickson, "What is machine learning data poisoning?", *TechTalks*, 7 October 2020. Available at: https://bdtechtalks.com/2020/10/07/machine-learning-data-poisoning/.

Other vulnerabilities include being subjects of "adversarial attacks", i.e., by finding a "a set of subtle changes to an input that would cause the target model to misclassify it. Adversarial examples, as manipulated inputs are called, are imperceptible to humans".[705]

Turner argues that it is "possible to distinguish between situations in which AI has made a mistake as to a fact and those in which AI has applied the 'wrong' rule to a known fact".[706] For example, "[w]hen a factory robot thinks that a human operator's head is a component in the manufacturing process and decides to crush it – killing the human – this is akin to a mistake of fact".[707] Conceivably, a case could be made in favor of raising an insanity defense if the robot mistook the human head due to a faulty neuroelectronic process (i.e., a mental process) caused by a virus (i.e., mental disease).

Let us look at an example made by Turner, which derives from the Holbrook v. Prodomax Automation Ltd case.[708] Wanda Holbrook was a journeyman maintenance technician working in an auto-parts factory in Michigan. Specifically, the company was deploying robots to manufacture trailer hitch receiver assemblies. The whole assembly line was run by a central computer running a software called programmable logic controller. Holbrook intervened on a robotic arm without following the employer-mandated safety protocols and was killed by a robot that crushed her head between the hitch assembly that was already in place. Another robot, then, tried to weld the new hitch assembly, severely burning her face, nose and mouth.

The Holbrook case, according to Turner, is different from the one of an AI supplanting human instruction in an unexpected way, such as "toaster burning a house down to cool all the bread",[709] and from the one of an AI developing the ability to deliberately

---

[705] Ibid.

[706] Turner, 2019, p. 204.

[707] Ibid.

[708] Holbrook's husband pursued a wrongful death suit, that is, a civil action, against six defendants, including the producer of the robot and the manufacturer who designed the automated assembly line. Before her death the opening of the door of the zone where one of the robots was would only cause that specific zone to power down, while raising walls to prevent robots in other zones from entering. After the accident the PLC was programmed to shut down every zone when a single door was open. According to his claim, the defendants were negligent in programming the software which ran the assembly line, since they had not programmed the PLC to stop every zone from the very beginning. The Michigan Western District Court qualified the PLC software as product under Michigan law, therefore dismissing the common law negligence claim brought by the plaintiff. *Holbrook v. Prodomax Automation Ltd.*, 1:17-cv-219 (W.D. Mich. Sep. 20, 2021).

[709] Turner, 2019, p. 204.

disobey clear human instructions". Arguably, these cases would resemble "a concept of criminally guilty mind", i.e., *mens rea*.

5.4.2 Artficial Infant Offenders

Let us move now to considering the second question, i.e., shall AI systems be treated as *infants* by the criminal justice system? If one looks at how infancy is regulated in a selection of states, it is possible to notice some common traits. Indeed, article 97 of the Italian penal code establishes a presumption of absolute lack of criminal capacity for minors under the age of 14. Nevertheless, according to articles 222 and 224 of the Penal code, a judge can still decide to adopt a number of coercive measures against the minor. Similarly, article 19 of the German StGB provides that a person under the age of 14 cannot be liable for criminal offenses – yet the liability of accomplices is not excluded (art. 29).

In the US, punishment against minors is under the jurisdiction of family law. Some US states have codified the common law infancy defense in their statutes, i.e., an absolute presumption of incapacity upon children under the age of 7.[710] Most states deal with the problem of juvenile offenders as a matter of establishing the age which triggers the jurisdiction of the juvenile court (usually 14). Few of these states proscribe a presumption of incapacity which can be rebutted in court by showing that the juvenile knew the wrongfulness of what she was doing.[711] Most of the times, juvenile courts do not place a minimum age for their jurisdiction: this entails that if a state has not incorporated the juvenile defense into juvenile law, admittedly a child under seven might be adjudged delinquent "for conduct for which *[she]* lacked criminal responsibility".[712]

As noticed by Hallevy, who asks himself if this reasoning could be applied to AI systems, the majority of AI systems are already "capable of analyzing permitted and forbidden".[713] Indeed, the category of infancy seems more fitting than the one of insanity, especially with regards to two aspects: first, the fact that there is often a hand out to other

---

[710] LaFave, 2017, p. 487.

[711] For example by proving that the individual knew what she was doing and that it was wrong, also based on the child's general knowledge of the difference between good and evil. See the judgments cited in LaFave, 2017, p. 486.

[712] LaFave, 2017, p. 488.

[713] Hallevy, "The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control", 2010, p. 190.

fields of law, which are deemed more adequate; second, with regards to the criminal liability of the (human/adult) accomplices.

The analogy between AI and children is often mentioned to point out to the drawbacks of state-of-the-art AI systems, which are "great at playing Go and playing chess and maybe even driving in some circumstances" but, at the same time "are terrible at doing the kinds of things that every two-year-old can do".[714]

Indeed, computer scientists are trying to create AI systems which are capable of doing "two-year-old style things",[715] specifically, that are capable of adopting the learning process of children.

In a nutshell, they are trying to generate AI systems able to do something that is different from what they are taught to do. Think of this analogy:

> […] if you look at the current models for A.I., it's like we're giving these A.I.'s hyper helicopter tiger moms.[716] There's a programmer who's hovering over the A.I. and saying, oh, yeah, yeah, you got that one right. That one's a cat. That one's a dog. That one's another cat. That one's another dog. Or you have the A.I. that's saying, oh, good, your Go score just went up, so do what you're doing there. But nope, now you lost that game, so figure out something else to do.
>
> And as you might expect, what you end up with is A.I. systems that are very, very good at doing the things that they were trained to do and not very good at all at doing something different. So they can play chess, but if you turn to a child and said, OK, we're just going to change the rules now so that instead of the knight moving this way, it moves another way, they'd be able to figure out how to adapt what they're doing. And it's much harder for A.I. systems to do that.[717]

---

[714] *Ezra Klein Interviews Alison Gopnik*, 2021.

[715] Ibid.

[716] Tiger and helicopter parenting are parenting styles. The first one comprises of parents "pushing their children to succeed according to their parents' terms", where the latter are those who "take over every aspect of the child's life". See: https://theconversation.com/from-tiger-to-free-range-parents-what-research-says-about-pros-and-cons-of-popular-parenting-styles-57986.

[717] R. English, "From tiger to free-range parents – what research says about pros and cons of popular parenting styles", *The Conversation*, 25 May 2016. Available at: https://theconversation.com/from-tiger-to-free-range-parents-what-research-says-about-pros-and-cons-of-popular-parenting-styles-57986.

Moreover, as stated by Diamantis et al., the analogy between algorithms and children wavers when it comes to identifying who should be liable for AI-harm:

> For each child … there are just one or two human beings who are *obvious candidates*. For algorithms, the picture is often more complex. Large, dispersed teams of humans work on today's most important algorithms. So, there's no obvious human to hold liable. But even if there were, it is not clear there would be any point to holding the human liable. These sophisticated algorithms are often beyond the capacity of any individual human to meaningfully influence.[718]

The solution that these authors propose is to hold corporations liable. This topic will be dealt with in at Para 6.4.2.

5.5 PRELIMINARY CONCLUSIONS

Let us draw some preliminary conclusions.

The issues addressed in this Chapter belong to the so-called general eligibility challenge theorized by Abbott and Sarch.[719] One can deconstruct the broad eligibility challenge into three components: assumptions (A) and (B) which lead to conclusion (C) ($A \wedge B \rightarrow C$). In these terms, the broad eligibility challenge holds that

A the capacity for culpable conduct, i.e., criminal capacity encoded in law in incapacity defenses (e.g., infancy and insanity) is a general perquisite of criminal law and "failing to meet it would remove the entity in question from the ambit of proper punishment; *and*

B AI lacks the practical reasoning capacities needed for being culpable; *therefore*

C AI does not fall within the scope of criminal law.[720]

---

[718] Diamantis, Cochran & Dam, 2022, p. 2 [emphasis added].

[719] See Para. 3.3.1.

[720] Abbott & Sarch, 2019, p. 350.

As it was already mentioned, these authors elaborate three possible solutions to the eligibility challenge: *respondeat superior*, strict liability, and a framework for direct *mens rea* analysis for AI. Let us focus now on the last one, specifically on what would entail for an AI system to be "culpable in its own right".[721]

Generally speaking, criminal law is indifferent towards unmanifested mental states or one's motives for breaking the law.[722] Rather, criminal justice systems are interested in those who express *insufficient regard* for protected legal goods. Indeed, one could argue that criminal law

> does not demand that we are motivated by respect for others, or even respect for law; all it demands is that we do not put our disrespect on display by acting in ways that are inconsistent with attaching proper weight to protected interests and values. Thus*, criminal culpability can be seen as being more about what one's behavior manifests* and less about the nuances of one's private motivations, thoughts, and feelings.[723]

It follows that, in this perspective, only the "legal" notion of culpability is relevant: as long as you "cross the line" and are capable of criminal conduct you are eligible for criminal punishments.[724] What matters is that one is "capable of behaving in ways that *manifest* insufficient regard for the legally recognized reasons".[725]

One can conceive criminal capacity of corporations building on this notion of culpability: since they "possess information-gathering, reasoning and decision-making procedures … [t]hrough their members, they weight and act on the reasons that criminal law demands not displaying insufficient regard for in action".[726] Hence, corporations can act in a way that "puts on display their insufficient regard for the legal interests of others".[727]

How does this translate to AI systems? Undeniably, AI systems also possess information-gathering, reasoning and decision-making processes: their behavior – which is

---

721 Ivi, p. 355.
722 A. Sarch, "Who cares what you think? Criminal Culpability and the Irrelevance of Unmanifested Mental States", *Law and Philosophy*, Vol. 36, No. 6, 2017, p. 708.
723 Abbott & Sarch, 2019, p. 356.
724 Abbott & Sarch, 2019, p. 356.
725 Ibid.
726 Ibid.
727 Abbott & Sarch, 2019, p. 357.

the result of a determination on which are "the most efficient means" [728] to reach their goals – could be deemed by the law functionally equivalent to manifesting insufficient regard for the contents of criminal norms, i.e., as criminal.[729] This argument echoes Hu's theory, who believes that the first requirement in order to impose criminal liability on robots is that they are equipped with moral algorithms.[730]

Moving forward, one can suppose that all criminal legal systems to some extent presume that citizens possess knowledge of the law and punish them accordingly. In fact, ignorance of the contents of the law does not exclude liability (*ignorantia legis non excusat*), except for cases of mistake of law defenses.[731] The same presumption works with regards to knowledge of moral rules:

> The law seemingly assumes that, beyond the knowledge of specific norms, individuals have reached a certain level of moral understanding. In this sense, the defendant must be at least able to procedurally follow the norms that are prescribed by the law. This means that she should be able to understand what the law requires and to modify her behaviour accordingly. To do this, the defendant must possess some basic requirements of rationality allowing her to convert general rules into everyday practices.[732]

Moreover, it can be assumed that criminal legal rules, due to their innate characteristics and functions of command, are easier to encode in the AI systems than rules of morality. In the case of AI systems, then, the presumption of knowledge of the law could be almost absolute. The same cannot be affirmed with regards to moral rules where, instead, the bar for AI systems is much lower than the one for human beings. In other words, we could soon be dealing with very legally abiding – yet very immoral – AI systems.

The heart of the issue, though, is slightly different: it is not an issue of whether machines are able to act as good citizens (i.e., morally right), it is more a matter of whether they can act wrongly out of their *own* free will. How do you encode that part of evil that

---

[728] Ibid.
[729] Ibid.
[730] See Para 3.2.2.
[731] Bonicalzi & Haggard, 2019, p. 109.
[732] Ibid.

maybe all of us possess and never express – the one that makes us capable to disregard a norm? How can you make that into a cost and benefit decision?

If one builds an AI system programming it (on a hypothetical level) to commit a crime, she will be passible of criminal liability; concomitantly, the AI system will possibly commit the offense only because it was given the possibility to do so by its creator. The fact that AI systems are human made products appears to be an unsurmountable issue. Nonetheless, it is not possible to draw definitive conclusions without looking at how such issues are surmounted in the field of the paramount man-made creation: corporate criminal liability.

To conclude, as mentioned in the beginning of this Chapter, personhood in criminal law is strictly tied to theories of criminal punishment: it is only in that context that it acquires meaning.[733] We must then ask ourselves: what would be the benefits of treating an AI system as an addressee of criminal law?

[733] Simmler & Markwalder, 2019, p. 21.

6 ARTIFICIAL INTELLIGENCE CRIME

6.1. Introduction: Glimmers of an Economic Theory of AI-Crime – **6.2.** Matters of Mens Rea – **6.2.1.** Overview – **6.2.2.** Responsibility of Machines – **6.2.3. –** Responsibility of humans - **6.2.3.1.** The DNA of Negligence – **6.2.3.2.** Human Oversight and False comforts – **6.2.3.3.** It's All About the Data – **6.3.** Matters of Actus Reus: Overview – **6.3.1.** The act – **6.3.2.** Failures of Causation – **6.3.2.1.** The 'Many Hands Problem' – **6.3.2.2.** The Black Box Problem – **6.3.2.3.** The Shortcuts Problem – **6.4.** Corporate Criminal Liability – **6.4.1.** Models of Corporate Criminal Liability – **6.4.2.** The Next Frontier? Diamantis' theory of Corporate Criminal Liability for Automated Decisions – **6.5.** Preliminary Conclusions

6.1 INTRODUCTION: GLIMMERS OF AN ECONOMIC THEORY OF AI-CRIME

Ch. 4.1. presented the PCM experiment.[734] We should draw four lessons from it. First, the PCM can be subsumed in the broader discussion on the AI alignment problem[735] and the creation of "moral machines", i.e., systems that are aligned with human (moral) values, which was discussed above at 4.1.

Second, as pointed out by Diamantis et al., the PMC "reveals a predicament":[736]

> Part of the reason AI can be so powerful is that it doesn't follow human commands. It uses massive data sets (more massive than any human could comprehend) to uncover complex solutions that no human could anticipate (or even understand). But by freeing AI from the constraints of low-level programming, AI like the PCM will inevitably also harm us in unforeseeable ways. AI's unpredictability is the source of both its power and its danger.[737]

---

[734] Diamantis, Cochran & Dam, 2022, p. 1.

[735] See, *inter alia,* I. Gabriel, "Artificial Intelligence, Values, and Alignment", *Minds and Machines*, Vol. 30, 2020, pp. 411-437.

[736] Diamantis, Cochran & Dam, 2022, p. 1.

[737] Ibid.

Third, we do not need Artificial General Intelligence (AGI)[738] to conceive an AI system which, as an "extremely powerful optimizer" could pursue goals "that are completely alien to ours", such as criminal ones, even if not as catastrophic as turning all humans into paperclips.[739]

Fourth, the key-role played by the concept of rational behavior. Why is rationality then relevant when discussing issues of criminal liability, specifically those related to AI systems? This aspect will be briefly discussed, using the PCM as reference. The PCM experiment is based on a *utility function*: produce the maximum possible number of paperclips.[740] Utility functions are one of the founding bricks of the (micro)economic theory of consumer choice.[741] According to this theory, individuals (consumers) are *rational* beings which make choices that will "yield them the highest level of well-being subject to the constraints that

---

[738] See the definition at p. vii.

[739] See Badea & Artus, 2022, pp. 1-2, who highlight that "[w]hat one can draw from the literature on ethics and machines, is that moral problems will be generated not by machines going rogue and deliberately trying to kill us (though this possibility cannot be ignored), but by machines inadvertently harming us while they try to carry out the instructions we have given them … it is not too far-fetched to think that in future we will be able to build machines that are such good means-ends reasoners, or goal maximisers that they will be able to think of creative new ways to achieve the ends that we provide for them".

[740] "… its utility function is linear in the number of paperclips times the number of seconds that each paperclip lasts, over the lifetime of the universe … the core premise is just that, given actions A and B where the paperclip maximizer has evaluated the consequences of both actions, the paperclip maximizer always prefers the action that it expects to lead to more paperclips". See "Paperclip maximizer", *Arbital*. Available at: https://arbital.com/p/paperclip_maximizer/.

[741] "Utility functions arise when we have constraints on agent behavior that prevent them from being visibly stupid in certain ways. … Suppose that you're a hospital administrator. You have $1.2 million to spend, and you have to allocate that on $500,000 to maintain the MRI machine, $400,000 for an anesthetic monitor, $20,000 for surgical tools, $1 million for a sick child's liver transplant … Should this hospital administrator spend $1 million on a liver for a sick child, or spend it on general hospital salaries, upkeep, administration, and so on? … if you cannot possibly rearrange the money that you spent to save more lives and you have limited money, then your behavior must be consistent with a particular dollar value on human life. By which I mean, not that you think that larger amounts of money are more important than human lives—by hypothesis, we can suppose that you do not care about money at all, except as a means to the end of saving lives—but that if you can't rearrange the money, then we must be able from the outside to say: 'Assign an X. It's not necessarily a unique X. For all the interventions that cost less than $X per life, we took all of those, and for all the interventions that cost more than $X per life, we [didn't take any] of those.'… The overall message here is that there is a set of qualitative behaviors and as long you do not engage in these qualitatively destructive behaviors, you will be behaving as if you have a utility function.". E. Yudkwosky, "The AI Alignment Problem: Why It's Hard, and Where to Start", Transcription of the speech given at Stanford University on May 5, 2016, p. 2. Available at: https://intelligence.org/files/AlignmentHardStart.pdf.

they face",[742] following a set of preferences expressed by a utility function. Utility functions, therefore, are used to describe "the level of well-being, or satisfaction, that the individuals gets from any combination of these two goods".[743]

AI systems are *extremely* rational agents compared to human beings.[744] At the same time, the concept of rationality is entangled with the one of good behavior. Indeed, "[a] rational agent is one that does the right thing".[745] Being a rational agent and doing the right thing entails, for each perceived sequence, selecting "an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has".[746]

What does rationality have to do with criminal liability? As it was already highlighted, criminal law assumes that an offender possessed some kind of rationality which led her to understand the command of a rule and to choose not to comply with it.[747] According to most, the *legally relevant* notion of rationality tied to the concept of criminal capacity comprises of a "general ability to recognize and be responsive to the good reasons that should guide action".[748] Therefore, people might "engage in legally relevant behavior for *nonrational*, *irrational*, and *foolish* reasons, but this does not excuse them or render them nonresponsible if they are generally capable of rationality".[749] As it was shown above, state-of-the-art AI systems do not fulfill such criteria, hence they could not be considered as *perpetrators* of crimes following a humancentric framework of criminal law.

---

742 T. J. Miceli, *The Economic Approach to Law*, Stanford University Press, 2004, Appendix to Chapter 1.
743 Ibid.
744 "An AI system is thus first and foremost rational". AI-HLEG, 2019, p. 2.
745 Norvig & Russel, 2003, p. 36.
746 Ivi, p. 37.
747 "Unless human beings are rational creatures who can understand the applicable rules and standards, and can conform to those legal requirements through intentional action, the law would be powerless to affect human behavior. Legally responsible agents are therefore people who have the general capacity to grasp and be guided by good reason in particular legal contexts. They must be capable of rational practical reasoning. The law presumes that adults are so capable and that the same rules may be applied to all people with this capacity". S. J. Morse, "Rationality and Responsibility", *Southern California Law Review*, Vol. 74, 2000, p. 353.
748 Ivi, p. 354.
749 Morse, 2000, p. 354.

Thus, let us look at a different conceptual framework. Enter: economic criminal theories.[750] This approach will not be tackled in depth, as it is not the purpose of this research. Nevertheless it is argued here that it could provide innovative insights which could inspire further research and, as such, it deserves to be addressed.

In his seminal paper "Crime and Punishment: An Economic Approach", Becker assumes that a person commits an offense "if the expected utility to him exceeds the utility he could get by using is time and other resources at other activities".[751] This entails that "some persons become 'criminals', therefore, not because their basic motivation differs from that of other persons, but because their benefits and costs differ".[752]

Since then, the debate on the economic analysis of criminal law thrived. It was argued that "one of the distinctive features of the economic analysis of criminal law—as compared to the traditional retributivist approaches to criminal law—is its focus on deterrence, on the social ends that are promoted through the imposition of punishment, rather than retribution and moral culpability".[753] According to Posner the economic approach to criminal law is a better positive theory of criminal law "since in so many areas conduct is punished that is not blameworthy in the moral sense".[754]

Let us transpose these concepts to AI systems: if we assume that the decision to commit a crime is the result of a cost-benefit analysis, we can ask ourselves whether this analysis could be carried out by an AI system. Preliminarily, one must note that AI systems would not incur in certain "costs" of crimes specifically tied to humans, such as the so-called "psychic costs", such as guilt, anxiety, fear, or shame, since they are not capable of feeling emotions. For the same reasons, AI systems would not be able to obtain "psychic benefits" from committing a crime, such as "the thrill of danger, peer approval, retribution, sense of accomplishment, or 'pure' satisfaction of wants".[755]

---

[750] The idea to approach this field of research came from presenting my research at the Institute for Transnational Legal Research (METRO), Maastricht University. I am thankful to prof. Niels Philipsen, who invited me to present my research, at and to prof. Michael Faure for his suggestion.

[751] G. Becker, "Crime and Punishment: An Economic Approach", *J. Pol. Econ,* Vol. 76, 1984, p. 178.

[752] Ibid.

[753] T. Fisher, "Economic Analysis of Criminal Law", in Dubber & Hörnle (Eds.), 2014, p. 42.

[754] R. A. Posner, "An Economic Theory of the Criminal Law", *Columbia Law Review,* Vol. 85, 1985.

[755] E. Erling, "Economics of Criminal Behavior", in B. Bouckaert & G. De Geest, *Encyclopedia of Law & Economics*, Elgar Publishing, 1997.

Indeed, if on the one hand concepts such as "moral culpability, fairness, justice, and retribution"[756] seem to malfunction when applied to AI systems, on the other, those of "rationality, utility maximization, or efficiency"[757] seem to work like a charm. The most interesting aspect, nevertheless, is that AI systems respond to costs and benefits of crime *better* than human beings because they are more *rational* agents in an economic sense: they possess (and can also gain through ML) more information about the environment that surrounds them and the consequences of their actions than human beings, hence they are capable of making more precise estimates of the probable benefits and costs of their criminal activity. This makes their choices innately more informed, hence more rational, than the ones of human criminal offenders.

To my own surprise, the fact that one could engineered obedience to obligations in an AI systems is an appealing statement. Indeed,

> […]artificial agents could respond to the threat of punishment by modifying their behavior, goals, and objectives appropriately. A realistic threat of punishment can be palpably weighed in the most mechanical of cost-benefit calculations. […] *An agent rational enough to understand and obey its legal obligations would be rational enough to modify its behavior so as to avoid punishment, at least where this punishment resulted in an outcome inimical to its ability to achieve its goals.*[758]

Let us assume that AI agents – in the future – will be capable to possess assets and funds and control money independently. They could be punished with financial sanctions, similarly as it is already done with corporations. In this sense, one could make a case for punishing AI systems following an economic approach to criminal law. Punishment could fulfil a deterrent function, provided that the expected punishment (price) that an AI system would face outweighs the expected benefits. A similar stance is expressed by Lagioia-Sartor[759] and Chesterman. Indeed, the latter argues that the economic analysis of punishment, "may seem particularly applicable to both corporations and AI systems".[760] The difference between

---

[756] Fisher, 2014, p. 40.
[757] Ibid.
[758] Chopra & White, 2011, pp.168-169 [emphasis added].
[759] See Para. 3.3.2.
[760] Chesterman, 2021, p 124.

AI systems and corporation is that when it comes to corporations incentives or deterrents are aimed at human subjects, where instead in the case of AI systems they would be aimed directly at its program – provided that it is architectured in a way as to "maximize economic gain without regard for the underlying criminal law itself".[761]

Moreover, Chesterman believes that an approach based on "narrowly tied penalties encouraging good behavior as well as discouraging bad",[762] similar to the one adopted with corporations, seems well suited for AI systems: in this sense, violations of the criminal law would be deemed as "errors to be debugged, rather than sins to be punished".[763] In other words, punishing AI systems could fulfill the *reform* purpose of punishment.

Reform aims at preventing future crime by changing the offender into a law-abiding citizen. In a utilitarian perspective, which considers that committing a crime means that "the wrongdoer has miscalculated the values of certain actions, or failed to consider the costs to others and only considered the benefits they might receive",[764] reform means to "revise people's utility functions".[765] This could be done, as mentioned above, through monetary sanctions : "in terms of economic decision making, the rational person will recognize the additional costs of getting caught and being punished, and thus avoid choosing illegal actions".[766]

Such an interpretation of reform rings a bell. Let us recall the functioning of a ML training method called *reinforcement learning*, whose working can be summarized as follows: "the idea is that if the robot makes the wrong choice and you want it to make the right choice in the future, you need to change its decision structure", typically by tweaking the probabilities of making a certain decision or changing the values placed on certain outcomes so that the decision process of the AI system is rewritten and the desired outcome becomes more desirable, hence more guaranteed in future behavior.[767] Some have even argued that

---

761 Ivi, p. 125.
762 Chesterman, 2021, p. 125.
763 Ibid.
764 P. M. Asaro, "Determinism, machine agency, and responsibility", *Politica & Società,* Vol. 2, 2014, p. 281.
765 Ibid.
766 Ibid.
767 Asaro, 2014, p. 282.

> punishing and fixing are exactly the same: punishing is a clumsy, external way of modifying the utility function. Furthermore, a closer analysis reveals that fixing or modifying the robot's utility function directly is tantamount to punishment, in the sense that the robot would not want it to happen and would act if possible to avoid it.[768]

Chesterman concludes by stating that "neither legal personality nor the coercive powers of the state should be necessary to ensure that machine learning leads to outputs that do not violate the criminal law".[769] To a greater extent, according to this opinion, AI agents could be "restrained by purely technical means, by being disabled, or banned from engaging in economically rewarding work for stipulated periods […] Particularly errant or malevolent agents (whether robots or software agents) could even be destroyed or forcibly modified under judicial order as dangerous dogs are destroyed by the authorities today".[770]

Finally, as mentioned in Chapter 3, some argue that punishing AI directly could fulfill a general deterrence aim of criminal law, not only towards other AI-systems ("punishment of an artificial agent … would nevertheless be educative for other artificial agents, given sufficient intelligence. After all, examples of corporate punishment are taken very seriously by other corporations")[771]; but also towards humans-in-the-loop, and have "expressive benefits" for victims of AI-harm.[772]

Thus, one must also underline that, as of now, it is unlikely that punishing AI systems through the means of criminal law would motivate law-abiding behavior from human citizens.[773] AI systems are not perceived as members of our society. In other terms, if we assume that one of the primary goals of criminal law is to "restitch" the social fabric breached

---

[768] J. Storrs Hall, "Towards Machine Agency: a Philosophical and Technological Roadmap", 2012, p.4. Available at: https://robots.law.miami.edu/wp-content/uploads/2012/01/Hall-MachineAgencyLong.pdf>.
[769] Chesterman, 2021, p 125.
[770] Chopra & White, 2011, p.167.
[771] Chopra & White, 2011, pp.168-169 [emphasis added].
[772] Abbott & Sarch, 2019, p. 345.
[773] Simmler & Marlwalder, 2019, p. 22.

by the wrongdoing,[774] modern expectations (of law-abiding behavior) of AI systems might not possess (yet) a sufficient "breaking character" to ask for criminal punishment.[775]

As argued by Seher,

> Even if criminal law dogmatics were to be based on a concept of action according to which non-human, computer-driven actions could be understood as violations of norms, this understanding would - so I claim - be limited to the insider circle of criminal legal scholars. For the foreseeable future, the unanimous opinion of the population is likely to be that they are "machines" or "computers".[776]

This issue will be analyzed further in the conclusions. For now, let us continue in the reflection: even if we agreed that a case could be made for punishing AI systems directly, AI-misbehavior would still need to fulfill the criteria proscribed for *mens rea* and *actus rea*. These two criteria will be analyzed in this order. The choice is not accidental: the first section, "matters of *mens rea"* will tackle separately intent-based responsibility of machines and negligence-based responsibility of humans. The section "matters of *actus reus*", instead, will present issues regarding to the objective element of the offense that are transversal and apply to both machines and humans.

6.2 MATTERS OF *MENS REA*

6.2.1 Overview

As it was highlighted in Ch. 3, the discussion in the field of AI and *mens rea* articulates in two directions. The first direction has to do with the possibility of conceiving of "guilty" AI systems. The second direction concerns the liability of the human agent from time to time

774 Simmler & Markwalder, 2019, p. 22.

775 Simmler & Markwalder, 2019, p. 26.

776 Translated from German: "*Selbst wenn aber die Strafrechtsdogmatik einen Handlungsbegriff zugrunde legte, nach dem nicht-menschliche, computergelenkte Aktionen als Normbrüche verstanden werden könnten, wäre dieses Verständnis doch – so behaupte ich – auf den Insiderzirkel der Strafjuristen beschränkt. In der Bevölkerung dürfte auf absehbare Zeit einhellig die Meinung vorherrschen, es handele sich um „Maschinen" oder „Computer*". Seher, 2016, p. 59.

involved, the so-called "*human-behind-the-machine*." This Chapter will deal with both topics in that order.

6.2.2 Responsibility of Machines

Chapter 3.2. highlighted how the supporters of a "robotic" *mens rea* can be counted on the fingers of one hand. This section will address the issue with an open mind, stripped – as much as possible – of any kind of anthropocentrism.

Think of the following example:

> a system of DNNs[777] is designed to devise a profitable trading strategy in the equities markets. It is given access to a broad range of data, including a Twitter account, real-time stock prices of thousands of securities, granular historical price data, and access to popular business news feeds. Within months of training on data, the algorithm is able to consistently turn a profit. It is unclear what strategy the AI has stumbled upon, but it is rapidly placing trading orders, consummating some of them and rapidly withdrawing or changing others. Interestingly, the system has learned to "retweet" news articles on Twitter and often does so before and after trades. The designer of the system is not able to tell what role the retweets have in the overall trading strategy, nor is he able to tell why certain trade orders are consummated and others withdrawn. All he can tell is that his AI is working and is profitable. Within days, the price of one of the securities that the AI frequently trades crashes steeply within seconds. The AI, which can either take a long or short position in the security, has, however, managed to make a profit. [778]

Could we say that the AI system committed the intent crime of market manipulation?[779] As a matter of fact, it was *not* designed to do so: its objective – by design –

---

[777] Deep Neural Networks, see definition at vi.

[778] Bathaee, 2018, p. 911.

[779] Market manipulation is a criminal offense in most legal systems and can be defined as the intentional artificial manipulation of the price of financial instruments (products, securities or commodities) through practices such as the spreading of false or misleading information and conducting trades in related instruments to profit from this. See, *inter alia*, the definition at art. 5 of the EU Directive on criminal

was the lawful activity of identifying which shares or stocks to buy/sell in order for the company to obtain a profit.

While AI-expansionists argued – more or less convincingly – that AI systems might display *mens rea* following the philosophical theory of BDI,[780] others have attempted at modelling a definition of criminal intent which could be suitable for algorithms.

In a seminal paper published in 2021, Hal Ashton presented a set of formal definitions of intent which could be converted into "fully formal language, fully suitable for an algorithm".[781] In order to create such a definition, one must assume that "the concept of intent exists outside the human mind",[782] similarly as it is done with corporations. As the author argues, his theory could be used to inform courts in the event of AI-crime "as to the culpability of its owner and programmer using the existing mechanism of secondary liability".[783] Hence, he did not write this paper with direct liability of AI systems in mind, but to find ways to exclude liability of human beings.

Ashton's theories are mentioned here, and not in Ch. 3, for two reasons: first, the author can be considered an "agnostic" in the debate ("This article is not going to make any claims about the eligibility of algorithms for legal personhood, blame, punishment or even praise and the role that algorithmic intent might play in that … The only thing that this article requires of the reader is that they are open to the possibility that intent (and related mens rea states) can exist in an algorithm")[784]; second, the author adopts an original approach which is more relevant to the discussion of this Chapter than to the literature review.

Intent presupposes that the "intended result must be foreseeable as a result of an act".[785] According to Ashton, ascertaining intent in algorithms might differ from humans, since machine-intent is "presumably perfectly observable (assuming some access to the algorithm)".[786] He contends that it is possible to "peer into" the algorithm and assess the

---

sanctions for market abuse (*Directive 2014/57/EU of the European Parliament and of the Council of 16 April 2014 on criminal sanctions for market abuse*).

[780] We will not address whether machines can display criminal negligence. Instead, we will solely focus on the highest mode of intent (i.e., direct intent or purpose). For this reason, from now on the word "intent" will be used to refer exclusively to *dolus*.

[781] Ashton, 2022, p.17.

[782] Ivi, p. 23.

[783] Ashton, 2022, p.32

[784] Ivi, p. 2.

[785] Ashton, 2022, p. 7.

[786] Ivi, p. 7.

"constituent parts" behind a definition of intent. In other words, he argues that whether an AI system has a certain judgment of likelihood regarding a certain result following an action is "a matter of observable fact".[787] At a first glance his statement seems to collide with the Black-Box Problem, which will be addressed further in this Chapter.

Ashton argues that, in order for an AI system[788] to "have the capacity to act with intention in a meaningful way", i.e., to possess criminal capacity, it should meet fundamentally two requirements. First, it should have "some sort of causal model of the world for it to be able to know whether action *a* has a causal relationship with variable X"[789]: by looking at this requirement, one can judge whether the AI system knows the consequences of its action. Second, it should have "some sort of preference ordering over states of the world",[790] by looking at this requirement, one can judge whether the AI system has an aim or a desire.

Assuming that these requirements are satisfied, Ashton proposes 3 definitions of intent. The first one is the most relevant for the purposes of this study:

Definition 1 (*Direct Intent at commission*) An agent D directly intends a result $X = x$ by performing action a if:

(DI1) **Free Agency** Alternative actions a' exist which D could have chosen instead of a.

(DI2) **Knowledge** D should be capable of observing or inferring result $X = x$

(DI3) **Foreseeable Causality** Actions a can foreseeably cause result x (according to D's current estimate).

(DI4) **Aim** D aims or desires result x.

Figure 10. Definitions of intent. Source: Ashton, 2022.

Ashton's definitions could be a useful tool mostly to evaluate liability of human agents, rather than the one of the machines. Let us as assume that, according to the definition provided above, the court were to ascertain that the AI system committed the market manipulation out of its own (unpredictable) "intent". Such information should be taken it

[787] Ashton, 2022, p. 7.
[788] Ashton speaks of "Agent".
[789] Ashton, 2022, p. 24.
[790] Ibid.

into consideration – as an excluding or mitigating factor – when evaluating the criminal liability of the managers of the corporation deploying the system, of the team that designed the system, etc.

6.2.3 Responsibility of Humans

Attributing direct liability to state-of-the-art AI systems proves to be a challenging task. We might inquire, then, whether subjects *other* than the AI system could be liable, jointly with the machine or as sole perpetrators. This research will not address cases where an AI system is *purposedly* used as a tool to commit a crime, as it does not represent a conceptually challenging scenario for the purposes of the general part of criminal law. Therefore, the following sections will exclusively deal with the criminal responsibility of humans for negligence-based offenses.

6.2.3.1 The DNA of Negligence

Although almost all contemporary legal systems acknowledge intention or purpose (*dolus directus*), classifications of the other types of guilty mental states vary. In particular, some systems, like the one outlined in the Model Penal Code, make a distinction between intention (the conscious desire to bring about the result), knowledge (of the prohibited result which will almost certainly follow the act), recklessness (doing an act while aware that it entails a substantial and unjustifiable risk of harm), and negligence (objective fault in creating an unreasonable risk).[791] Accordingly, recklessness is distinguished from negligence based on the element of awareness of the risk: a defendant might be considered *negligent* when she was not aware of risk, but she should have been, where instead a defendant might be considered *reckless* when she was aware of the risk and consciously disregarded it. Therefore, this classification stipulates that the distinction between negligence and recklessness is not the risk that is created, which is the same (i.e., a substantial and unjustifiable risk), but rather that the reckless actor *consciously* ignores the risk, whereas the negligent actor does not.

In continental legal systems, such as the German and the Italian one, the distinction between different kinds of *mens rea* is less fine-grained,[792] as they discriminate only between intent (*dolus*) and negligence (*culpa*). As a result of this bipartite scheme of (guilty) mental

[791] LaFave, 2017, p. 243.
[792] J. Blomsma & D. Roef, "Forms and Aspects of Mens Rea", in Keiler & Roef (Eds.), 2019, p. 179.

states,[793] negligence includes also other "intermediate forms" of subjective responsibility, such as recklessness.[794] In these systems negligence represents "the most normative form of mens rea" and is "primarily based upon a violation of the required duty of care"[795] which then results in a prohibited outcome.

Negligence also requires an "individualizing standard" for the reasonable person,[796] i.e., the assessment of negligence is connected to the individual skills of the defendant.

Having acknowledged this, it is not necessary for the purposes of this analysis to adopt a stance in this debate. What is important to recall is that some forms of criminal liability, to which we will refer to with "negligence" as an umbrella term – are based on based on two elements: (a) foreseeability and (b) breach of a duty of reasonable care.[797]

---

793 Although further internal distinctions can be identified.

794 T. Weigend, 'Subjective Elements of Criminal Liability', in Dubber & Hörnle (Eds.), 2014, p. 498. The issue then becomes how to qualify the conduct of "conscious risk taking": the *escamotage* can be found in the *dolus eventualis* and conscious negligence doctrines. *Dolus eventualis* can be described as a conduct of intentional risk taking: 'the actor does not know whether his conduct will bring about a harmful result but accepts the occurrence of that result "in the event that" it comes about'. In other words, the agent 'mentally embraces that outcome'. Conscious negligence, instead, can be described as a conduct of negligent risk taking: the actors do not know whether their conduct will bring about a harmful result, in fact, they unreasonably reject the idea or do not take this possibility seriously (a sort of 'everything will be alright' kind of mental state), but still decide to take the risk. Some argue that the distinguishing element between the two should not be the volition element, rather the knowledge one: the real difference between dolus eventualis and conscious negligence (or *luxuria*), then, lies in whether the agent knew that there was a grave risk of harm or a minor risk of harm. See also G. Fletcher, "'The Theory of Criminal Negligence: a Comparative Analysis", *University of Pennsylvania Law Review*, Vol. 119, 1971.

795 Blomsma & David Roef, 2019, p. 195.

796 N. Jan, "Autonomous weapons systems: new frameworks for individual responsibility", in Bhuta N. et al. (Eds), *Autonomous weapons systems. Law, ethics, policy*, Cambridge University Press, 2016, p. 12.

797 If we focus on the US legal system and on the MPC a person is regarded as acting *recklessly* (section 2.0.2(2)(c) a)): "When he consciously disregards a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that, considering the nature and purpose of the actor's conduct and the circumstances known to him, its disregard involves a gross deviation from, the standard of conduct that a law-abiding person would observe in the actor's situation". Section 2.02(2)(d) establishes that a person is acting negligently: "When he should be aware of a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of his conduct and the circumstances known to him, involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation". American Law Institute, *Model Penal Code: Official Draft and Explanatory Notes, C*omplete Text of Model Penal Code as Adopted at the 1962 Annual Meeting of the American Law Institute at Washington, D.C., 24 May 1962.

Negligence is generally based on the detachment from a "golden" standard of conduct. Furthermore, negligence has a "normative essence",[798] and comprises of an *objective* and a *subjective* dimension. The latter consists in the enforceability of the respect of such a duty of care: the agent must be reproachable, as he should have and could have avoided the harmful outcome. This means that the law could have demanded from him a better conduct. Such requirement is satisfied when the event was foreseeable and concretely avoidable by the model-agent, based on a number of factors such as the specific hazardous activity at stake and on the individual qualities of the agent. As it was argued, predictability and awareness of the risk are the DNA of negligence.[799]

The former consists in the conduct of violation of a pre-existing norm establishing the duty of care. The agent must have had – more or less strongly – *foreseen* and – more or less strongly – *accepted the risk* that an unlawful consequence will arise from the conduct. The norms may be codified or not codified and have the purpose of pre-emptively striking a balance between conflicting values: on the one hand, the benefits of a certain dangerous activity; on the other, the goods which are endangered by the activity.[800] Moreover, the harmful event caused must be avoidable through the observations of these norms. In other words, the alternative correct conduct must be suitable to avoid harm.[801]

According to Diamantis et al., negligence in the field of humans "behind" the AI systems could arise in different scenarios: "selecting improper data sets to train an algorithm, unreflectively specifying its success conditions, insufficiently testing it before release, or even inadequately hardcoding prohibitions for the algorithm ('No matter what, do not turn people into paperclips.')".[802] Let us take a closer look at these issues.

6.2.3.2 Human Oversight and Human in The Loop: Begging the Question?

As it was already exposed, one of the issues connected to AI systems that leads to a vacuum in the allocation of liability is its autonomy. How should this gap be filled?

---

798 Mantovani, 2020, p. 358.
799 Piergallini, 2020, p. 1765.
800 Mantovani, 2020, p. 363.
801 Ivi, p. 369.
802 Diamantis, Cochran & Dam, 2022, p. 7.

If we turn the gaze to the AI Act, it seems that one of the solutions to this problem should be found in the concept of "human oversight" and the use of so-called "Human-In-The-Loop" (HITL) techniques.[803] These consist in the creation of AI systems in which the model (*output*) is developed through interaction with a human agent who, for example, can play the role of "teacher" in the training phase of the system, providing *feedback* to the machine on the result obtained.[804] Tentatively, "if an autonomous system causes harm to human beings, having a human in the loop provides trust that somebody would bare the consequence of such mistakes".[805]

Such solutions, however, present some problematic issues. In a recent article titled "the false comfort of human oversight as an antidote to AI harm",[806] authors Ben Green and Ambra Kak believe argue that placing humans back in the "loop" of AI seems reassuring, but it is actually loopy in a different sense: it rests on a circular logic that offers false comfort and distracts from inherently harmful uses of automated systems. Policymakers, in fact, are turning to humans to mitigate the risks posed by AI systems based on the assumption that humans are indeed able to police their decision-making processes. In other words, human oversight seems to be used as kind of "band-aid" on the issues posed to AI autonomy.

Beck contends that when a human in the loop is included in the decision-making process "one has to realise … that in many situations, this might lead to excessive demand and responsibility of the human in question".[807] In the contest of AVs, for example, the driver requires at least 6 seconds to overtake control on the vehicle, which is in most traffic situations not enough to avoid incidents.

It might even be the case that instead of an actual oversight on the action of the machine the human is only capable of "rubber-stamping" the results produced.[808] The risk is

---

803 European Commission, *Artificial Intelligence Act,* art. 14.

804 For a survey of HITL approaches applicable to *machine learning*, see X. Wu et al., "A Survey of Human-in-the-loop for Machine Learning", *arXiv:2108.00941*, p. 2021.

805 European Parliamentary Research Service, "The ethics of artificial intelligence: Issues and initiatives", 2020, p. 35. See also, Pizzi, Romanoff & Engelhardt, 2020, pp. 145–180; AI-HLEG, 2019.

806 B. Green & A. Kak, "The False Comfort of Human Oversight as an Antidote to A.I. Harm Human Agency in Decision-Making Systems", *Slate*, 15 June 2021.

807 Beck, 2018, p. 43.

808 B. Wagner, "Human Agency in Decision-Making Systems", *Policy & Internet*, Vol. 11, No.1, 2019, p. 114.

to create a (human) scapegoat: the contrast with common principles of criminal law is self-evident.

How will a human, in practice, be able to supervise a system that was created to overcome it and make up for her shortcomings? This is especially relevant if translated in the field of negligent liability.

6.2.3.3 Negligence Failures

Simplifying matters to the extreme, as it was stated above, one can affirm that the essential elements of culpability comprise the failure to observe cautionary rules of conduct (codified or non-codified) and the assessment of the concrete agent's behavior in comparison to a model agent. In addition, these precautionary rules have a preventive purpose: the harmful event or danger constitutes the realization of the risk that they were created to avoid.

Negligence failures can be defined as "situations in which the classical building blocks of negligence, ie, risk taking, foreseeability, and awareness, struggle to identify a liable human being to whom we can attribute AI-caused harm. One could envisage negligence failures as nothing but a further development of the 'irreducibility challenge', first theorized by Abbott and Sarch, applied specifically in the field of criminal negligence". [809]

Now, in order to transpose these concepts into the topic under scrutiny, we first need to acknowledge that AI systems are already deployed in "permissible risk" activities (such as transport) which are allowed by our society. Furthermore, as argued by Guerra et al., "[a]t the level of utmost generality, it is important to bear in mind that human negligence and machine error do not represent equivalent risks … The social cost of machine error promises to be drastically lower than that of human negligence" [810].

It follows that AI systems operate in areas where there are already rules of conduct in place. However – and it is here that one can grasp the peculiarity related to AI – it is necessary to ask whether or not our society currently possesses the correct technical-scientific knowledge to apply pre-existing rules of conduct specifically to the new scenarios that these

[809] A. Giannini, J. H. Kwik, "Negligence Failures and Negligence Fixes. A Comparative Analysis of Criminal Regulation of AI And Autonomous Vehicles", *Criminal Law Forum*, 2023, p. 3.
[810] Guerra, Parisi & Pi, "Liability for robots I", 2021, p. 1. The authors reason in the field of tort wrongs, nevertheless their arguments can be transposed and analyzed to criminal liability.

systems will cause, or even to develop new ones. The answer to this question would seem, for the time being, to be in the negative.

In other words, although one might consider the invocation of the precautionary principle by European authorities to be appropriate, I agree with those who believe that the precautionary principle should serve as a beacon for legislative policy choices, rather than for penalizing ones.[811] This does not preclude the possibility that in the future the threshold required to impose the adoption of new precautionary rules will be exceeded, thanks in part to progress in scientific laws, particularly those that will be concerned with investigating, reconstructing and explaining the behavior of the most complex AI systems.[812]

In a slightly different vein, some scholars argue in favour of creating a new legal concept, namely "fault or negligence by programming", which would thus shift the responsibility for the harm committed by the "creature" onto the "creator".[813] The following questions then arise: how should this kind of negligence be shaped? More specifically, how should the precautionary rules of conduct be characterized in the implementation and deployment of an AI system?

Subsequently, one can point out certain specific critical issues pertaining to the precipitous aspect of the *enforceability* of the dutiful conduct – omitted by the "human supervisor" – which can be traced back to the broader discussion on the concept of "human in the loop" as exposed earlier.

In particular, one of the most problematic junctures consists of the *level* of care that can be demanded of the potential supervisor, be it the driver of the semiautonomous self-driving car or the physician using an AI-based diagnostic system. Consider, for example, *automation complacency*, a term coined in the field of aviation accidents to refer to the phenomenon whereby the automation of any task leads the human supervisor to trust that the machine is handling it effectively and, as a result, to stop paying attention;[814] or *automation*

---

811 F. Giunta speaks in terms of a "*criterio di politica legislativa*". F. Giunta, "Il diritto penale e le suggestioni del principio di precauzione", *Criminalia*, 2006, p. 229.

812 C. Brusco, "Rischio e pericolo, rischio consentito e principio di precauzione. la c.d. "flessibilizzazione delle categorie del reato", *Criminalia*, 2012, p. 389.

813 Magro, 2014, p. 516; Manes, 2020, p. 4.

814 L. Smiley, "'I am the Operator': The Aftermath of a Self-Driving Tragedy", *WIRED,* 8 March 2022. Available at: https://www.wired.com/story/uber-self-driving-car-fatal-crash/. One of the earliest definitions of *automation complacency* is that developed by E.L. Wiener, "Complacency: Is the term useful for air safety?", *Proceedings of the 26th Corporate Aviation Safety Seminar*, 1981. More recent ones include R.

bias, which is the tendency of human beings to place undue trust in the recommendations produced by a computer system.[815] These phenomena, amplified in the case of AI-based automation, would appear to significantly reduce the attention threshold of the human agent grappling with the machine, and, as a result, decrease the threshold of the diligent conduct demandable from the 'model' agent.

What is more, researchers recently have introduced the theory of the "big red button", a sort of emergency kill switch for AI systems which could stop it before it becomes destructive.[816] Such mechanisms need to be handled with care, especially if they are to be translated to criminal liability. Should the HITL be liable for "not pushing the kill switch"? At which conditions? Indeed, "accountability mechanisms built on the assumption of a supreme human overseer are inherently flawed, if adopted without criticism. Such approaches can embed and reinforce the implicit human/machine dichotomy and mystify human agency".[817]

On this matter, Gaede argues that subjects such as "the passenger", "the owner", or the "manufacturer of an AV" could be in principle punished for disregarding standards of care. Moreover, he believes that the most pressing issues are the standard of care that would be impose on these subjects and how responsibility would be shared amongst them.[818] For example, one could investigate whether negligence could be attributed to a "researcher" who unintentionally – but in violation of defined standards of care related to AI – creates or releases into the market a dangerous AI system. Gaede contends that, in case of a violation of these future duties (which would be tailored to the development of AI systems), it would not be possible to excuse the researcher based on the claim that she lacked foreseeability of

---

Parasuraman & D. H. Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration", *Human Factors,* Volume 52, No. 3, 2010.

815 It is defined as the "tendency to use automation as a substitute heuristic for vigilant information seeking and processing" by K. Mosier & L.J. Skitka, "Automation use and automation bias", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1999, p. 344.

816 Arnold & Scheutz, 2018, p. 60.

817 R. Koulou, "*Human Control over Automation: EU Policy and AI Ethics", EJLS*, Vol. 12, No. 1, 2020, pp. 9-46.

818 ""*[d]ie drängenden Fragen liegen hier darin, welche Sorgfaltsstandards wir den Beteiligten nochauferlegen wollen und wie sich Verantwortungsanteile zueinander Verhalten*". K. Gaede, "Künstliche Intelligenz – Rechte und Strafen für Roboter? Plädoyer für eine Regulierung künstlicher Intelligenz jenseits ihrer reinen Anwendung", in E. Hilgendorf & S. Beck (Eds), *Robotik und Recht*, Vol. 18, Nomos, 2018, p. 81.

the damage, since nowadays foreseeability does not require a detailed foresight of the final harmful event.[819]

Indeed, the trend to depart from a strict definition of foreseeability is visible in criminal law efforts to regulate *risk* and in the application of negligence constructs given by courts. In other words, what is asked of the addressee of the criminal norm is not the foreseeability of the *concrete harmful event*, but of a *class* or *type* of events and even of *worst-case-scenario*s.[820] In Italy this change is visible starting from the decisions on asbestos exposure cases up to the ones in the field of natural catastrophic events such as the Vajont Dam[821] disaster or the 2012 earthquake in L'Aquila.[822]

The analogy between AI-harm and natural disaster is also evoked by Jacob Turner. Turner, quoting John Danaher,[823] admits that we are facing a retribution gap which is caused by the "delta between humanity's expectations that someone will be held responsible, and our present inability to apply criminal law to AI".[824] In his opinion there are two feasible options: treating the actions of AI as "Acts of God" which have no legal consequences, or finding a "responsible" human. Thus, he believes that "[u]nlike earthquakes or floods, the acts of AI are unlikely to be viewed as unfortunate but morally neutral natural disasters".[825]

On the one hand, attaching criminal liability to programmers or users might lead to over deterrence and therefore slow down innovation. On the other, attaching criminal liability to AI systems could in theory fill the retribution gap, even though it would be difficult

---

819 "*Wurden zukünftig geltende Sorgfaltsregeln der KI-Forschung nicht geachtet, könnte ihn das Argument der mangelnden Vorhersehbarkeit zumindest nicht allgemein entlasten, da die Vorhersehbarkeit schon heute keine detaillierte Voraussicht des letztlich verletzenden Geschehens verlangt. Der Forscher weiß darum, dass sich die Folgen der selbst lernenden Technik schwer eingrenzen lassen bzw. muss er dieses Wissen in seine Betrachtungen instellen".* Gaede, 2018, p. 81.

820 "*La rivoluzione copernicana (o galileiana) posta in atto dalla giurisprudenza di merito e poi di legittimità consiste nell'assumere a 'parametro' della prevedibilità, ai fini del giudizio di colpa, non già il fatto storico hic et nunc accaduto, bensì un 'genere', una 'classe' di eventi, nei quali il fatto concreto risulti sussumibile: così facendo, l'evento viene ridescritto per così dire "a tavolino" quale species di un genus, con conseguente espansione dell'area di prevedibilità e (dunque) di responsabilità*", Civello, *La "colpa eventuale" nella società del rischio. Epistemologia dell'incertezza e "verità soggettiva" della colpa*", 2013, p. 118.

821 Cassazione Penale, Sez. IV, 25 March 1971.

822 Cassazione Penale, Sez. IV, 25 March 2016, No. 12748.

823 J. Danaher, "Robots, Law and the Retribution Gap", *Ethics and Information Technology*, Vol. 18, No. 4, 2016.

824 Turner, 2019, p. 120.

825 Ibid.

to locate intent in the system. Assuming that the system's mental state could be "measured and ascertained", the question which remains then is the one of whether "it would be appropriate from a social and psychological perspective to apply criminal law tenets to a non-human entity" or if it were better to define new mental states which would be applicable to AI and with a label different from "mens rea".[826]

If such an abstract concept of foreseeability of harm is deemed sufficient for the purposes of criminal law, it becomes then apparent how arduous it would be for the relevant human agent to deny that she had not abstractly foreseen the potential risks associated of deploying, for example, ML techniques.[827] Think of the case of an engineer who created an AI system for making toast and imagine that this system then burns down a house, causing the death of the people who lived in it, because it followed the reasoning "all the bread would be toasted". [828] The programmer, then, might be criminally liable in case it were proven for his "reckless behavior in creating such a program".[829]

Gless, Silverman, and Weigend contend that the criminal negligence bar should be lower when it comes to AI applications that generally reduce the risk of harm to society, according to an application-by application approach (rather than trying to find "all-or-nothing" rules).[830] Such an approach is commendable: the standard of care should be modelled according to the social value of the AI-application. E.g., "standards of care should be stricter with respect to robots that are of lesser social value, such as toys. With respect to self-driving cars, on the other hand, the risk remaining after careful testing and monitoring may be offset against the general benefits of using such cars".[831]

Provided that there will be areas of application (such as self-driving cars or medical devices) in which it is statistically proven that the use of AI system reduces the risk of harm, compared to when the same activities are performed by humans, let us imagine two scenarios for the future. Scenario A: the humans behind the machine exercise the greatest care possible. Nevertheless, the AI system commits harm which falls into the sphere of a "generally

826 Turner, 2019, p. 204.
827 Magro, 2020, p. 20.
828 Turner, 2019, p. 119.
829 Ibid.
830 Gless, Silverman & Weigend, 2016, p.430.
831 Ivi, p. 436.

foreseeable malfunctioning".[832] This act should be regarded as a "normal risk".[833] Scenario B: the humans behind the machine exercise the greatest care possible. Nevertheless, the AI system commits harm that falls outside the reasonable and foreseeable sphere of risk connected to its area of application. This act should be regarded as an event interrupting the chain of attribution,[834] same as a lighting hitting a tree (the so-called "coincidental intervening cause").[835]

Both scenarios, then, should lead to an exclusion of criminal liability. In other words, we should consider them as cases of "just bad luck". In this sense, this researcher agrees with Lima's standpoint:

> […] humans should learn to live with this unfortunate development, much in the same vein that they have learned to live with the results of a bridge collapsing due to a hurricane or a flat tire that leads to a car accident. *Not everything can be foreseen, prevented, or contained, and in everyday life there are several instances where no one is to blame-much more be held criminally liable-for an undesirable outcome.* In other words, *not everything can or should be regulated under criminal law*. Depending on the familiarity that humans will develop with AI agents in the future, this option might prove to be a viable alternative to criminal liability, even though policy implications have to be considered as it is likely that AI acceptance rates might suffer at first.[836]

The job of defining what falls into the sphere of "acceptable risk" and what falls out, if no rules of conduct or standards are available, is one for the judges. After all, this is what is done already by courts and it appears as there are not sufficient reasons justifying a change when it comes to crimes committed "by" AI systems. The resulting "gap" in accountability should be filled by other domains of law, such as torts or administrative liability. Criminal law shall not be regarded as a *panacea* for all evil, not even for robo-evil.

---

832 Ivi, p. 433.
833 Ivi, p. 432.
834 Gless, Silverman & Weigend, 2016, p.432.
835 Ibid.
836 Lima, 2018, p. 694.

6.2.3.4 It's All about the Data

A question that remains unanswered is the following: who exactly is the human operator?

When analyzing the ample literature on criminal law and AI,[837] together with policy papers on "ethical AI",[838] two things catch the reader's eye. First, little or no attention is given to the different roles which incur in the development chain of an AI systems. Most papers refer only to the programmer, who then becomes the epicentre of accountability. However, AI teams comprise of different figures, such as data analysts and scientists, who take care of data collection and interpretation, and ensure that collected data is relevant and exhaustive while also interpreting the analytics results; data, and machine learning engineers, who build and test machine learning models; and then also programmers, who implement the solution, transforming the model into coding. This is relevant for various reasons. Most importantly because without a correct understanding of who-does-what, one cannot allocate liability in a way that is compatible with principles of criminal law, notably with the principle of blameworthiness. Second, there is little or no reference to liability connected to mistakes or inadequacies of training data. Why are these two factors important for criminal liability?

Let us consider a real-world scenario. An algorithm based on ML techniques is capable of teaching itself rules by learning from the training data through statistical analysis, detecting patterns in large amounts of information and generating outputs. The patterns can then be applied in different tasks, such as driving a car.[839] Deep learning, a subset of ML, is a technique which classifies information through layers of artificial neural networks (ANNs).[840] The networks have input nodes, which are fed raw data, and output nodes, which determine to which category the input information belongs. For example, this system might be used to classify whether an image contains a woman or not. There are many layers between the input and the output one: in order to obtain an output, the algorithm needs to be trained with large sets of labelled data (for examples, pictures labelled as containing a woman). ML algorithms are hungry for data: the more classified data they are fed, the more accurate will be their prediction. The issue is that it is impossible to provide every possible labelled sample

---

837 See Ch. 3.

838 See Ch. 7.

839 Surden, 2019, p. 1311.

840 B. Dickson, "What are artificial neural networks (ANN)?", *TechTalks*, 5 August 2019. Available at: https://bdtechtalks.com/2019/08/05/what-is-artificial-neural-network-ann/.

of data to a deep learning algorithm. In other words, you cannot feed the algorithm with all the pictures of animals, or persons, or cars in the world. Consequently, the algorithm will have to generalize between its examples to classify data that it has never seen before. What happens, then, when the algorithm has been trained on faulty, biased, or wrong data? Indeed, one could mention the now (in)famous Uber's self-driving car fatal accident as an example.[841] This short digression allows us to grasp the importance of good quality training data and of who's behind it.

Indeed, datasets are "rife with errors".[842] At the same time, machine learning (ML) is being used in critical settings, where a mistake could prove fatal. One could cite as an example the now famous fatal accident of Uber's self-driving car,[843] caused (also) by the inability of the vehicle's software to identify the presence of a pedestrian crossing on the road without using a crosswalk.[844]

Recently, data and machine learning engineers started to address the issue. For example, the "data centric AI"[845] (DCAI) movement aims at shifting the focus of ML engineering from modelling to the underlying data used to train and then evaluate models. Aspects such as the collection and generation of data, the labelling of data and the evaluation of its quality are central to avoid errors, and therefore harm. Yet, they have not gained popularity in relevant criminal doctrine.

6.3 MATTERS OF *ACTUS REUS*

[841] BBC, "Uber's self-driving operator charged over fatal crash", 16 September 2020,. Available at: https://www.bbc.com/news/technology-54175359.

[842] D. Kang et al., "Finding Errors in Perception Data With Learned Observation Assertions", *Stanford Dawn*, 24 January 2022. Available at: https://dawn.cs.stanford.edu/2022/01/24/loa/.

[843] D. Wakabayashi "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam", *The New York Times online*, 19 March 2018. Available at: https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

[844] "The system never classified her as a pedestrian-or correctly predicted her path-because she was crossing N. Mill Avenue at a location without a crosswalk, and the system design did not include consideration for jaywalking pedestrians". See National Transportation Safety Board (NTSB), *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018* ("NTSB Report Tempe"), Highway Accident Report NTSB/HAR-19/03, p. 16. Available at: https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf.

[845] See https://datacentricai.org.

6.3.1 The Act

As highlighted in the introduction, *actus reus* is a concept which includes both questions of conduct and of causation. Focusing on the first, in order to have liability there must be a *bodily* movement.[846] One can define "act" in a broad way, thus including involuntary acts as sleepwalking, or in a narrow way, as a "willed" bodily movement. [847]

If one were to adopt a broad interpretation of "act", as predicated by Hallevy and Lagioia/Sartor, it would follow that an AI system could act in a criminally relevant way, i.e., that it could perform the *actus reus*. In other words, if one were to adopt a materialistic notion of action, "AI systems – not only as robotics applications but also as algorithmic systems of pure software – might be considered capable of realizing a criminal action, as realized in external relationships, both in cyberspace and in the "physical" world".[848]

If, on the other hand, one were to adopt narrower interpretation, it would follow that in order for a conduct to be willed, i.e., to establish the link between mind and body, the agent would have to display "consciousness" and "control" of the act.[849] In Anglo-American legal doctrine, for example, it has been argued that punishing an act without a voluntary component would not fulfill functions of criminal law such as deterrence and retribution, while achieving restraint or rehabilitation (even though it would "probably [*be*] best to deal with this outside the criminal law").[850]

Can the act of an AI system fulfil a strictly-construed act requirement of a criminal offense? Corporations, as a matter of fact, are treated – in certain legal systems – as subjects of criminal law which possess "rational agent capacities".[851] Similar to AI agents, they can take "rational decisions on the market and possess the capacity for self-organisation and to adapt themselves to the ever-changing economic environment … Thus, corporations although lacking feelings and emotions nevertheless possess capacities for intelligent agency".[852] This is deemed enough to make them susceptible of criminal liability (provided

---

[846] We will not consider in this section crimes consisting of an omission to act.
[847] LaFave, 2017, p. 304.
[848] AIDP General Resolution Section I, 2022, pp. 25-26.
[849] Keiler, 2013, p. 61.
[850] LaFave, 2017, p. 305. See also the MPC § 2.01 providing that " a person is not guilty of an offence unless his liability is based on conduct that includes a voluntary act".
[851] Keiler, 2013, p. 61.
[852] Ibid.

that the other elements of the criminal offense are satisfied) and might also be sufficient for AI systems.

6.3.2 Failures of Causation

As already highlighted, "failures of causation"[853] is an effective expression coined by Ugo Pagallo to the describe the fact that AI agents break down the classic cause and effect analysis linked to matters of legal causation. There are multiple factors which could lead to a failure of causation. This Chapter identifies three of them: the 'many hands problem', the black box problem and shortcuts.

The factors which lead to failures of causation are not mutually exclusive: they co-exist, and their combination further muddies the web of causation by triggering mainly two consequences: (1) they challenge the identification of the (legally relevant) cause that led to the realization of the adverse event (*locus of liability issue*); (2) assuming that all the (legally relevant) causes can be identified, they challenge the identification of the degree of relevance that the single factor had on the realization of the adverse event (*weight issue*).

Finally, one must acknowledge that Human-Computer interaction,[854] in the forms of human oversight and HITL approaches, and the interaction between AI systems, also might have an impact in matters of legal causation. Human oversight and HITL were already addressed above. Interacting AI systems entails the fact that algorithms might cooperate "in a complex and dynamic ecosystem",[855] therefore leading to an exponential growth of failures of causation.

6.3.2.1 The "Many Hands Problem"

Let us start with the 'Many Hands Problem'.[856] It can be defined as the simultaneous of a multiplicity of actors involved in the conception, production, marketing, and use of AI

---

853 Pagallo, 2013, p.73.

854 Council of Europe & Yeung, 2019, p. 64.

855 Ivi, p. 67.

856 The phenomenon was first analyzed in the field of moral philosophy by D. F. Thompson, "Designing Responsibility: The Problem of Many Hands in Complex Organizations", *The American Political Science Review,* Vol. 74, No. 4, 1980, p.9. Subsequently, the development of the literature has been exponential. See, by way of example, M. Bovens, *The quest for responsibility. Accountability and citizenship in complex*

systems, as well as to the compartmentalization of the roles of individual human agents throughout this process and of their responsibilities. AI systems are the final product of individual contributions by "multiple individuals, organisations, machine components, software algorithms and human users, often in complex and dynamic environments".[857]

To understand this phenomenon, the example developed by F. Santoni de Sio and G. Mecacci might come in handy:

> a vehicle may be operated by a driver D1, with the assistance of the automated driving system AS, produced by the car manufacturer X, powered with digital systems developed by the company Y, possibly including some form of machine learning developed by the company Z, and enriched by data coming from different sources, including the driving experience of drivers D2, D3...Dn; vehicles in this system are in principle subject to a standardization process done by the agency S, the traffic is regulated by the governmental agency G, drivers are trained and licensed by the agency L etc.[858]

The issue engrained in the 'Many Hands Problem' is that, in contexts which are characterized by complex organizational dynamics such as the one under scrutiny, legal constructs struggle to identify the subject (human or non-human) which misbehaved and to reconstruct the hierarchy of powers that is behind the decision-making processes. Thus, theories of corporate criminal liability might come in handy on this matter.

6.3.2.2 The Black Box Problem

The Black Box Problem can be defined as "an inability to fully understand an AI's decision-making process and the inability to predict the AI's decisions or outputs".[859] Indeed, "it may be impossible to tell how an AI that has internalized massive amounts of data is making its

organizations, Cambridge University Press, 1998; I. R. Poel et al., *Moral Responsibility and the Problem of Many Hands*, Routledge, 2015.

[857] Council of Europe & Yeung, 2019, p. 11.

[858] F. Santoni de Sio & G. Mecacci, "Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them", *Philosophy & Technology,* Vol. 34, 2021, p. 1062.

[859] Bathaee, 2018, p. 905.

decisions",[860] same as it may be impossible to tell how a human brain functions.[861] In other words, you might be able to put a human being on the stand and question her, or she might leave a paper-trail of evidence behind her, but the same cannot be done with an AI system.[862]

Think for example to DNNs:[863] they consist of thousands or hundreds of thousands of neurons which, together, deliver an output (a decision) such as identifying the subject of a photograph. Even if one of the layers or clusters might "encode some feature extracted from the data, (e.g., an eye or an arm …)" often what happens is that "what is encoded will not be intelligible to human beings".[864] Moreover, the network learns from its experience through an intuitive decision-making process and such a process, most of the times, will not be reducible to a set of instructions "nor can one in most cases point to any neuron or group of neurons to determine what the system found interesting or important".[865] If we go back to the vehicle example mentioned above, its performance might be "potentially re-designed by the second on the basis of new data acquisition and processing - and opaque, if the reasoning scheme underlying systems' actions is not easily accessible to their controllers, regulators, or even their designers".[866]

To conclude, the issue of Black Box has obvious repercussions on the scope of negligence liability for the "human-behind-the-machine" discussed above. Not only it renders an etiological investigation on the realization of harm impractical from a *condicio sine qua non* perspective, it also does not place a "reasonable" person in the position to apply any precautionary measure or "risk calculus" to the – since she does not know what the risk is.

6.3.2.3 The Shortcuts Problem

Shortcut is a phenomenon that occurs in ML, specifically in self-supervised ML. Self-supervised ML is a model which learns by itself: it is fed raw data without any label description from humans, and it is then told, for example, to classify this data. For example, it might be fed hundred thousand images and being told to classify images which contain a

860 Ivi, p. 891.
861 Bathaee, 2018, p. 891.
862 Ivi, p. 892.
863 Deep Neural Networks, see vi.
864 Bathaee, 2018, p. 902.
865 Ivi, p. 903.
866 Santoni de Sio & Mecacci, 2021, p. 1062

cow.[867] Sometimes, it might happen that a model relies on a simple characteristic in a dataset (for example, the green grass) rather than learning the true meaning of that data. This phenomenon is referred to as a "shortcut". The result is inaccurate predictions. So, in this example the model might learn to classify an image of cows not by focusing on the shapes and the patterns of the cow, but by focusing on the green grass.

Supervised models are especially useful in for medical imaging. For example, a self-supervised learning model might be trained to classify whether a chest X ray represents a case of pneumonia or not. Imagine that the model now classifies an image, tags it as a pneumonia case, not based on the white spots on the lungs of the patient, but based on which hospital the scan comes from, or even worse, based on the gender or the nationality of the patient. This was the case studied in research published in May 2021,[868] where researchers examined models that had been deployed to detect covid-19 from chest X rays. They claim that rather than learning medical pathology, the systems identified what are called spurious associations, in other terms, they relied on shortcut learning to identify associations between medically irrelevant factors and a disease status. The system appeared accurate when applied in one hospital but failed to be as accurate when applied to new hospitals.

Let us imagine now that a doctor based his diagnosis on the result of an AI system and, since there was an error in the result provided by the system caused by a shortcut, the misdiagnosis results in the death of a patient. The following question, then, would be regarding how a judge shall apply the standard rule on attribution, that is, the operation through which he selects the legally relevant cause of a harmful event, in cases where the AI system takes a shortcut. What if it is not possible to reconstruct the web of causation *ex-post*? Which scientific theory should the judge rely on in order to ascertain liability?

6.3.2.4 Omissions to Act

Notwithstanding that most of criminal offenses punish active conducts, criminal law might also be extended to punish failures to act, even when the criminal offense is formulated only in active terms (i.e., requiring an active conduct with causes a result). Thus, in order for so-

[867] A. Zewe, "Avoiding shortcut solutions in artificial intelligence', *MIT news*, 2 November 2021. Available at: https://news.mit.edu/2021/shortcut-artificial-intelligence-1102.

[868] A. J. DeGrave, J. D. Janizek & S. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal", *Nature Machine Intelligence*, Vol. 3, 2021.

called conducts of commission-by-omission to be criminalized, there must be a legal duty to act.

Having acknowledged this, it becomes necessary to conduct a brief reflection on the possibility of identifying a duty to act (specifically, a "duty to control conduct of others")[869] upon the human overseeing the act of the AI systems. According to some authors, it is possible to identify a duty to act in the specific field of semi-autonomous cars[870] upon the driver present inside the vehicle[871] which would be "activated" when the driving system requires the driver to regain control of the car.[872]

Nevertheless, at this time the issue does not appear significant for a number of reasons. Among them is the absence of a fundamental aspect of perpetration-by-omission: as of today, there is no specific *legal* duty to prevent an AI system from causing harm, even though the general rules of criminal law on the matter continue to apply. Neither European, nor international, or domestic legislation contain express provisions in this regard. Yet, legal duties to act can arise also from other sources, for example upon the creation of an imminent danger. The danger in question could be the one caused by an AI system. Hence, one should question whether on it would be possible to subsume AI systems within the applicative sphere of the already existing legal duties to act by way of interpretation.

Undoubtedly, from a *de iure condendo* perspective, any effort in this direction would have to be suitable to identify *exactly* who the subject invested with such a duty is, provided that she must also possess an *effective* power of impediment, i.e., she must be physically capable to perform the act which would avoid the realization of the harmful event, in accordance with the founding principles of criminal law.

---

869 LaFave, 2017, p. 316.

870 This corresponds to level 3 of the SAE J3016 standards, i.e., vehicles that can put in place all the maneuvers necessary for driving the vehicle (e.g., acceleration braking). The person inside the vehicle, therefore, does not for all intents and purposes drive it, even if she is sitting in the driver's seat. She will only have to intervene if an explicit request to do so is made by the AI system. This type of technology is called "*hands and feet free but not 'mind free' driving*" by V.A. Banks et al., "Subsystems on the road to full vehicle automation: hands and feet free but not 'mind' free driving", in *Safety Science*, Vol. 62, 2014, pp. 505-514.

871 This would be, in particular, a "duty to control". Cf. Piergallini, 2020, p. 1751.

872 See in this regard Manes, who argues "[…] *è chiaro che se il guidatore non è richiesto di monitorare il traffico sino a una richiesta di riassunzione della funzione di guida, lo stesso non può più essere ritenuto in concreto* 'human in command' *né dunque (penalmente) responsabile di eventuali causazioni lesive occorse sino a quel momento, ove appunto il controllo sulla attività rischiosa era delegato alla macchina* AI driven *legalmente autorizzata*". Manes, 2020, p. 4.

The introduction of an "all-encompassing" duty to act, i.e., entailing an extended obligation for an unidentified "human overseer" to prevent any harm caused by an AI system, would be in conflict with the principle of legality, specifically with the principle of "strict construction of criminal statutes"[873] (*principio di tassatività*).

Finally, when one opens up to commission-by-omission, it must also confront herself with the enormous difficulties of ascertaining causation which is innately tied to omission cases. As it was already underlined, when dealing with the actions of AI systems one is confronted with the simultaneous presence of a myriad of alternative causal factors, both human and non-human. This makes it impractical, on the one hand, to identify the single factor that has not been activated to prevent or interrupt the causal process that has already begun and, on the other hand, to exclude alternative causal factors with the certainty required by modern criminal legal systems

6.4 CORPORATE CRIMINAL LIABILITY FOR AUTOMATED DECISIONS

6.4.1 Models of Corporate Criminal Liability ("CCL")

Before focusing on Diamantis' theory on "algorithmic corporate misconduct", it is relevant to deliver a brief overview of CCL, focusing on the most important aspects for this research.

According to some, "the real reasons for the existence of CCL could in fact lie in the power structures behind corporations. By punishing corporations one can prevent the shareholders, owners, and managers from having to go to prison and receiving moral blame, as society has found someone else to be blamed-the entity of a corporation".[874]

Approaches to regulating the criminal liability of legal persons can be divided into two major macro-families.

On the one hand, there are the systems that hold that corporations are merely a legal fiction, nothing more than the sum of the actions of multiple human figures (so-called nominalistic approach).[875] According to this model, the liability of the corporation derives

[873] J. L. Corsi, "An Argument for Strict Legality in International Criminal Law", *Georgetown Journal of International Law*, Vol. 49, 2018, p. 133; LaFave, 2017, p. 88.
[874] S. Beck, "Mediating the Different Concepts of Corporate Criminal Liability in England and Germany", *German L.J.,* No. 11, 2010, p. 1110.
[875] Keiler & Roef (2019), 336.

not only from the individual responsibility of the individuals who represent "the heart and mind"[876] of the legal entity, but also from those who are its "feet and hands"[877]: the top figures (such as managers and members of the executive) and individual employees. Such a responsibility is based on the doctrine of vicarious liability (under the principle *respondeat superior*), according to which the corporation is held responsible for the actions of the individuals who acts on its behalf, just as an individual, such as an employer, can be held responsible for the actions of its employee.

On the other hand, there are the systems that instead consider the corporation as a social reality in its own right, additional to the sum of the actions of its individual members. According to this model (so-called organisational approach), corporations are therefore capable of fulfilling the subjective (*mens rea*) and objective (*actus reus*) requirements of the offense independently of the individual human contribution and in ways that differ from it.[878] According to some, punishing corporate wrongdoing could even lead to the abandonment of the *mens rea-actus reus* dichotomy in favor of a single notion of corporate blameworthiness. In fact, as Roef notes, the two models are not mutually exclusive, and this is attested to by the fact that some jurisdictions have developed a mixed approach to corporate criminal liability. A third path exists, namely that of corporate administrative liability: this is the model adopted by Italy and Germany.[879] It is the legacy of a conception of criminal law strongly based on the principle of culpability and the preventive function of punishment.

The Italian model is an interesting case study. Corporations are not *criminally* liable since their liability is defined as administrative. Nevertheless, their liability is ascertained during a criminal trial; they are subject to the same guarantees and principles ruling criminal liability of natural persons; a *mens rea* connection between the fact and the corporation must be established in order to establish their liability. Interestingly so, the liability of the corporation and the one of the employee are independent (see art. 8 leg. decree 231/2001): it follows that a corporation might be responsible even if it was not possible to identify the specific employee which committed the offense or when the employee cannot be punished, e.g., because she lacked criminal capacity due to insanity.

---

876 Keiler & Roef (2019),336
877 Keiler & Roef (2019), 336
878 Keiler, 2013, pp. 437 ff.
879 Keiler & Roef (2019), 336

Germany's (administrative) corporate liability was established by the (*Gesetz über Ordnungswidrigkeiten* (OWiG) in 1975. The legislative project of introducing criminal liability, through the Verbandssanktionengesetz (VerSanG) failed in 2021. According to Dubber, it is too simplistic to state that "the general take on corporate criminal liability in Germany is that it does not exist, could not exist, and—not surprisingly—did not exist".[880] CCL existed in Germany but it no longer does: "The story then becomes not that there never was or could have been corporate criminal liability in German law, but that there never was or could have been corporate criminal liability in *modern* German law". [881]

Conversely, the United States represent the poster child of legal systems that provide direct criminal liability for corporations. These, in fact, can be prosecuted at both federal and state level. It has been affirmed that the United States pursue CCL tenaciously.[882] Admittedly, one of the reasons to pursue CCL is that it "increase[s] incentives for corporations to monitor and prevent illegal employee conduct",[883] i.e., that the threat of sanctions on corporations will induce them to "take steps to prevent the illegal conduct in the first instance, thus reducing the risk that employees will offend".[884] In other words, *respondeat superior* would be justified under a deterrence rationale.

In this respect, one must highlight the very wide scope of application of the American *respondeat superior* doctrine: corporations may be deemed liable for an employee's actions and mental state, committed within her employment with the intention of benefitting the corporation (at least in part), even if they were in violation of corporate policy.[885]

---

880 M. Dubber, "The Comparative History and Theory of Corporate Criminal Liability*", New Criminal Law Review: An International and Interdisciplinary Journal* , Vol. 16, No. 2, 2013, p. 204.

881 According to a simplistic account the reason would be "[c]orporate criminal liability in a modern, enlightened science of criminal law is illogical, impossible, unthinkable because it flies in the face of one of that science's greatest discoveries, if not its single greatest achievement: the "guilt principle" (*Schuldgrundsatz*, Latinized *ex post* as *nulla poena sine culpa*)". Markus Dubber, The Comparative History and Theory of Corporate Criminal Liability, New Criminal Law Review: An International and Interdisciplinary Journal , Vol. 16, No. 2 (Spring 2013), 205.

882 K. E. Goodpaster, "Tenacity: The American Pursuit of Corporate Responsibility", *BUS. & SOC'Y REV.*, Vol. 118, 2013, pp. 577-605.

883 United States v. Sun-Diamond Growers of Cal., 138 F.3d 961, 971 (D.C. Cir. 1998), aff'd, 526 U.S. 398, 1999.

884 R. Luskin, "Caring About Corporate 'Due Care': Why Criminal Respondeat Superior Liability Outreaches Its Justification", *American Criminal Law Review*, Vol. 57 2020, p. 303. See also E. Tuttle, "Reexamining the Vicarious Criminal Liability of Corporations for the Willful Crimes of Their Employees", *Clev. St. L. Rev.*, Vol. 70, 2021, p. 121.

885 Luskin, 2020, p. 312.

Particular emphasis is placed on corporate culture as a factor in determining the severity of the sanction response and the applicability of parole. In addition, an important factor is that of prosecutorial discretion: in most cases, corporations are able to avoid indictment if they follow specific steps, such as modifying their organizational structure, or paying fines, thanks to Deferred and Non-Prosecution Agreements.[886] Corporate culture is a determining facture also in quasi-criminal corporate liability systems as the Italian one: according to article 6 of Leg. Decree 231/2001, corporations are not liable for crimes committed by their high-ranking employees in case they prove, amongst other things, the adoption of an efficient organisational and compliance mode.

One could ask herself whether AI systems could be treated as corporate agents who committed a crime for "corporate reasons". According to a model of CCL such as the American one, it would not be required to establish whether the AI systems could fulfill the elements of a criminal offense in order to establish the corporation's liability.

The American approach differs from the English one for example, which is instead based on the identification doctrine: the *mens rea* of the top individuals, that is, those who can be considered the directing minds and wills of the corporation, are elevated to represent the subjective element of the entity itself.[887] In other words, this doctrine is based on an imputation mechanism such that the *mens rea* of the individuals who "personify" the corporation is attributed directly to the corporation itself. There is no unambiguous definition of what level of authority an agent must possess in order for it to be identified with the corporation. As has been argued, "[b]y requiring that only the most senior persons can be the 'directing mind and will' of a company, it's arguable that large companies are let off the hook, since many key decisions will be decentralized away from the most senior management."[888] Under this model, therefore, the legal entity is regarded as a social reality distinct from the individuals who work for it.

What impact does the existence of CCL have on liability of AI systems? As already mentioned, according to some there is no substantial difference between CCL liability and

886 United States Attorneys' Manual § 9-16.325, *Plea Agreements, Deferred Prosecution Agreements, Non-Prosecution Agreements and Extraordinary Restitution*, 2008.

887 A. Shalchi, House of Commons Library, Research Briefing, "Corporate criminal liability in England and Wales", CBP 9027, 2022, p. 5.

888 Shalchi, 2022, p. 8.

AI criminal liability.[889] The analogy appears cunning at times. Many questions arise. Surely, one would first have to establish what an AI system is made of: are AI systems made of algorithms, of mathematical operations, or are they constituted of the human agents who created them and control them? In the last case, would it mean that AI engineers, data scientists, etc., constitute the AI entity? What effect would then be punishing the AI system have on these subjects following a deterrent rationale? Surely, it could have powerful indirect effects on these subjects "behind-the-machine". Indeed, "criticizing or disparaging an AI agent may motivate its maker to change the agent's design".[890] In this sense, it can be argued that since AI agents are (or were, at some point) operated by humans, these humans could be deterred from the punishment of the AI systems, specifically in cases where the AI system *benefits* them substantially (be it in economic terms, such as a corporation deploying an AI system to make investment, or in emotional terms, such as robotic caregivers). Nevertheless, "criticizing the operation of a device is quite different from holding the device morally responsible".[891]

6.4.2 The Next Frontier? Diamantis' theory of Corporate Criminal Liability for Automated Decisions

The PCM thought experiment, according to Diamantis et al., snubs a key player: the corporation "which designed, owns and runs the PCM", and profits from it, "even if that entails converting some humans (preferably not customers!) into raw materials".[892]

Starting from the assumption that "the law is not equipped to address corporate liability when the 'thinking' behind corporate misconduct has been offloaded to automated systems",[893] Diamantis first developed a doctrinal framework "for extending the corporate mind to the algorithms that are increasingly integral to corporate thought".[894] According to this theory, any system carrying out the same functional role of an employee, including an

---

889 Hallevy, "The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control", 2010, p. 201.

890 V. R. Bhargava & M. Velasquez, "Corporate Responsibility And Artificial Intelligence", *The Georgetown Journal of Law & Public Policy*, 2019, p. 838.

891 Ivi, p. 837.

892 Diamantis, Cochran & Dam, 2022, p. 1.

893 Diamantis, "The Extended Corporate Mind: When Corporations Use AI to Break the Law", 2020, p. 898.

894 Ivi, p. 893.

AI system, can be part of the corporate mind.[895] As a consequence, it can be argued that corporations possessed "culpable knowledge"[896] of the algorithmic misbehavior.

The work of Diamantis then continues to argue that "algorithmic action is corporate action",[897] i.e., corporations can act both through their employees and through their algorithms. This would not require recognizing algorithms as independent beings. Algorithms would qualify as part of the "body corporate"[898] since they are entities upon which a corporation can exercise substantial control on and from which the corporation gains "substantial productive benefits".[899]

Diamantis declines the "Many Hands Problems" in a "corporate perspective":

> While there is usually at least one corporation behind most important algorithms, there are often many. One corporation may have designed a module for an algorithm that a second assembled. A third corporation may have tested the algorithm. A fourth may have marketed it to a fifth that owned and licensed it to a sixth that operated it on hardware owned by a seventh. Any doctrine for holding corporations vicariously liable for algorithmic harms must say in any circumstance which of these is on the hook.[900]

In one of his most recent works, Diamantis et al.[901] evaluate whether three current models of corporate liability (the purpose model, the strict liability model, and the negligence model) would be suitable to address algorithmic harm according to whether they can fulfill four goals:

(1) identify which corporation will be liable, i.e., make sure that the rules on liability can identify accountable corporations in a clear manner;

---

895 Diamantis, "The Extended Corporate Mind: When Corporations Use AI to Break the Law", 2020, p. 917.
896 Ivi, p. 920.
897 M. E. Diamantis, "Algorithms acting badly: A Solution from Corporate Law", *GEO. Wash. L. Rev.*, Vol. 89, 2021, p. 809.
898 Ivi, p. 829.
899 Diamantis, "Algorithms acting badly: A Solution from Corporate Law", 2021, p. 843.
900 Diamantis, Cochran & Dam, 2022, p. 4.
901 Ivi, p. 3.

(2) avoid gamesmanship, i.e., the liability model should not be easily manipulable by corporations, for example via loopholes;
(3) generate efficient incentives, i.e., the model should strike an efficient balance between too much or too little liability in order to guarantee social gains and innovation;
(4) produce fair outcomes, i.e., striking a balance between the rights of victims not to bear all the costs of algorithmic harms and those of defendants be unfairly prosecuted.

The *purpose model* – applying the *respondeat superior* theory – would occasion that the corporation would be liable for the harm committed by the AI system *purposely* designed or used by an employee to commit harm. Such a model would fulfill goals 1 and 2 but it wouldn't fulfil goal 3 and 4 since it fails to recognize accidental (and, one may add, negligent) accidents. The *strict liability model*, specifically the one based on products liability, could apply to algorithmic harms only in cases where the corporate algorithms qualify as products and the victims as consumers, which are not many. For example, in cases of injury caused by a self-driving vehicle, it would be easier for the purchaser of the car to qualify as a consumer rather than for the injured pedestrian.[902] The authors discard a broader strict liability model which would require a corporation, e.g., the one who owns the system, to be liable for any offense committed by the AI system, due to the fact that it wouldn't fulfill the aforementioned four goals.[903] Finally, the *negligence model*, which comprises of punishing corporations when its employees negligently designed or used an AI system which then caused harm, would fulfil goals 1 and 2, while leaving goals 3 and 4 dissatisfied. This model would impose "too much liability to be fair to corporations" and "too little liability to be efficient". Accordingly, AI systems could cause "serious but preventable harms, even if there's no way to prove that any human involved with the algorithm was negligent" for two reasons: the many hands problem and the "No Hands Problem".[904]

---

[902] Diamantis, Cochran & Dam, 2022*,* p. 6.

[903] "Unlike with strict products liability (where there's just one possible defendant—the manufacturer), it's not clear under a sweeping strict liability approach which corporation in the long chain of development of an algorithm should be held liable when the algorithm hurts someone (Goal 1). Any easy answer—like the corporation that owns the algorithm or the corporation that operates it—opens itself to easy manipulation (Goal 2). In holding corporations maximally liable for algorithmic harms, a broad strict liability model might overly depress corporate investment in algorithms (Goal 3). Lastly, since it requires no evidence of corporate fault, the strict liability model will inevitably punish innocent corporations, even if they did everything within their power to design and deploy their algorithms responsibly (Goal 4)". Diamantis, Cochran & Dam, 2022*,* pp. 7-8.

[904] Ivi, p. 8.

Focusing on the latter, it entails that:

> Each individual's contribution to a complex joint effort may be too miniscule for any one of them to count as negligently causing a harmful outcome. While the many hands problem was about the difficulty of finding evidence of negligence, the no hands problem is more metaphysical. A group of people can cause harmful outcomes, even if no one in the group was at fault. Under these circumstances, it is legally impossible to hold the corporation for an algorithmic harm because there is no negligent employee.[905]

Moreover, "[c]orporations know about the no hands problem. So they have an incentive to parcel out responsibilities among many different employees as a strategy for blocking their liability should something go wrong".[906] The authors conclude arguing for a "new model of corporate liability, tailor-made for algorithms", such as requiring "all corporations that use AI to pay annual dues into a public victims fund".[907]The last landing of Diamantis' theory is the one of "employed algorithms". In his latest publication, Diamantis argues that the relationship between corporations and algorithms should be thought as a type of employment.[908]

To conclude, the approach put forth by Diamantis is commendable as it adopts a "minimally invasive method"[909] to solve a soon-to-be very invasive problem. Nevertheless, we need to question whether this solution would be applicable also in countries whose legal system are not characterized by a "socially entrenched"[910] secular law of corporate liability which makes it "politically bulletproof",[911] such as the US. An element in favor of the general applicability of Diamantis' theory can be found in a recent Italian publication,[912] where the

---

905 Diamantis, Cochran & Dam, 2022, p.8.
906 Ibid.
907 Diamantis, Cochran & Dam, 2022, p. 9.
908 Diamantis, 2022.
909 Diamantis, "The Extended Corporate Mind: When Corporations Use AI to Break the Law", 2020, p. 901.
910 Ivi, p. 903
911 Diamantis, "The Extended Corporate Mind: When Corporations Use AI to Break the Law", 2020, p. 903.
912 F. Consulich, "*Flash offenders*. Le prospettive di *accountability* penale nel contrasto alle intelligenze artificiali devianti", *Rivista Italiana di Diritto e Procedura Penale*, Vol. 3, 2022, p. 1041 ff.

author introduces a variant of the *respondeat superior* model which would comprise of a transposition of the conduct from the AI agent to a different subject (be it an individual or the corporation). According to this theory, it would be possible to establish the (exclusive and autonomous) liability of a corporation for an offense committed by an AI system deployed in a complex organization. Assuming that the AI system can only be considered as a tool to commit a crime, the author argues that it would be possible to hold the corporation liable even in those cases where it is impossible to establish a subjective connection between the AI harmful conduct and a human agent. In other words, the conduct of the human agent, involved in the corporation, would work only to establish an objective nexus between the corporation and the act of the AI system. Accordingly, the corporation could be deemed liable for having deployed a maleficent AI system or for not having exercised the necessary safety controls (c.d. "*colpa di organizzazione*"). This would be in compliance with art. 8 of legislative decree 231/2001, titled "Autonomy of the corporation's responsibilities", which proscribes that the liability of the corporation occurs even when (a) the perpetrator of the crime has not been identified or cannot be charged; and (b) the crime is extinguished by a cause other than amnesty.[913]

6.5 PRELIMINARY CONCLUSIONS

The purpose of this Chapter was to analyze, with "non-judgmental" eyes whether it would be feasible to apply a traditional criminal legal framework (comprising of criminal capacity, *mens rea* and *actus reus*) to AI systems. Probably, this "legal imagination" exercise makes us, at least, moderates. Indeed, if one was to take a pragmatic stance to criminalization, we could be persuaded that holding AI systems directly liable, and therefore subject to criminal punishment – whether through fines, reprogramming or deletion, would fulfil a deterrence function.

Thus, criminal law does not function in a vacuum. A 2021 research on "how people's moral judgments of automated systems may clash with existing legal doctrines" showed suggest a conflict between people's desire to punish AI and robots and the punishment's

913 "*Art. 8. Autonomia delle responsabilità dell'ente 1. La responsabilità dell'ente sussiste anche quando: a) l'autore del reato non è stato identificato o non è imputabile; b) il reato si estingue per una causa diversa dall'amnistia.*" D.lgs 31/2001.

perceived effectiveness in achieving deterrence and retribution.[914] Indeed, "people wish to punish AI and robots even though they believe that doing so would not be successful, nor are they willing to make it legally viable".[915] Thus, the participants in the study believed that robots could learn from their mistakes, hence that their punishment could fulfil a pedagogic function.

As much as one would like to, it is not possible to forget about criminal law's intrinsic punitive nature and its retributive function: on this matter, this researcher's heart beats with the skeptics. This is not to say that criminal legal systems should pursue retribution at all costs: it is merely argued here that since criminal law cannot ignore its retributive nature, and the retributive function of criminal law cannot be fulfilled by agents that cannot display blameworthiness (such as AI systems), it follows that *criminal law is not the answer*.

In other words, holding AI systems directly liable for criminal offenses would not fulfill the so-called "retribution" gap.[916] Previous Chapters addressed the issue of "liability" and "responsibility" gaps. A clarification on the term "responsibility" is perhaps needed at this point.

Responsibility refers to a relationship between "an agent, its actions, and the outcomes of those actions".[917] In point of fact, as contended by R.A. Duff, responsibility is a "relational" concept, since it concerns the relationship that is established between a responsible person (A), an object (X) for which this person is responsible, and a third party (B). Responsibility for Duff thus means *"answerability,"* the mechanism by virtue of which A *is held* accountable for X by, and to, B.[918]

This relationship can take many forms:

> The first is that of *causal responsibility*, which denotes a causal link between the agent, their actions, and some particular outcome. The second is that of *moral/legal*

---

[914] Lima et al., 2021.

[915] Ivi, p. 6.

[916] Danaher, 2016.

[917] Ivi, p. 300.

[918] R. A. Duff, "Who is Responsible for What, to Whom?", *Ohio State Journal of Criminal Law*, Vol. 2, 2005, p. 442. See also R.A. Duff, "Moral and Criminal Responsibility: Answering and Refusing to Answer", in J. Coates & N.A. Tognazzini, *Oxford Studies in Agency and Responsibility Volume 5: Themes from the Philosophy of Gary Watson*, 2019, pp. 165-190; R.A. Duff, "Legal and Moral Responsibility", Philosophy Compass, Vol. 4, Issue 6, 2009.

> *responsibility*, which denotes the fact that the causal link between the agent and the action/outcome is such that the agent is an appropriate subject of legal/moral blame. This is usually determined by whether the agent has the right capacities and whether those capacities were exercised at the relevant time. The third is *liability responsibility*, which denotes the punishments or sanctions that an agent must bear in virtue of its moral/legal responsibility.[919]

The third denotation ("*liability responsibility"*) can be further differentiated depending on the relevant area of law. In the case of criminal law, it can be defined in terms of "punitive-liability" and "and is about suffering harm and public condemnation for wrongs done".[920]

Accordingly, the "retribution gap" arises due to the fact that one might – in light of what was underlined in the paragraphs above – consider that an AI system could bear both causal responsibility and moral/legal responsibility (the first and second denotations of responsibility), but not "liability responsibility" (the third denotation). In other words, it is argued that AI systems give rise to unsurmountable "gaps associated with the attribution of retributive blame for wrongdoing".[921] Said gaps cannot be ignored: human beings are, by nature, "inclined to punish in accordance with retributive criteria".[922]

Probably, as it is done with infants, and should be done with insane offenders, the solution lies in "looking elsewhere". In the field of torts law, for example, Guerra et al. argue for the introduction of a new tort liability regime called "manufacturer residual liability" ('MRL') which would apply to robots operated with human intervention and would shift "liability to manufacturers provided that operators and third-party victims have invested in due care".[923] In other words, manufacturers would be "strictly liable when operators and victims are not negligent regardless of design or manufacturing defects".[924] By doing so, it would accomplish four objectives: "(1) efficient levels of human care by operators and victims, (2) efficient activity levels in the use of robots; (3) efficient R&D investments for

---

[919] Danaher, 2016, p. 300.
[920] Ibid.
[921] Danaher, 2016, p. 301.
[922] Ivi, p. 303.
[923] A. Guerra, F. Parisi & D. Pi, "Liability for robots II: an economic analysis", *Journal of Institutional Economics*, Vol. 18, Issue 4, 2021, p. 1.
[924] Ivi, p. 2.

the development of safer robots; and (4) adoption of safer robots in the marketplace".[925] This issue will be dealt in greater detail with in the conclusive Chapter.

[925] Guerra, Parisi & Pi, "Liability for robots II", 2021, p. 2.

7 OVERVIEW OF EXISTING LEGAL FRAMEWORKS ON AI CRIMINAL LIABILITY

Laws, like sausages, cease to inspire respect in proportion as we know how they are made.

(Quote wrongly attributed to Otto Von Bismarck which presumably origins from John Godfrey Saxe, quoted on The Daily Cleveland Herald, Mar. 29, 1869)

7.1 Introduction – **7.2** General-Scope Tools – **A** Council of Europe – **B** Singapore: – **7.3** Self-Driving Tragedies: AV-specific tools – **C** England and Scotland – **D** France: – **E** Germany – **7.4** Conclusions

7.1 INTRODUCTION

Upon initiation of this study in 2019, I decided to maintain a stance of impartiality, and probably of hope, on whether legislation specific to AI in the realm of criminal law would emerge. Consequently, this Chapter's existence was established from the beginning. It chapter makes use of examples of legislation across the world and, as such, it is not a compete recollection. At the end of my study (i.e., early 2023), only few countries had developed legislation specifically dedicated to AI and Criminal Law.[926]

Before discussing these samples of regulation, it is relevant to make some general remarks on the topic of regulating AI.

[926] See, e.g., the General Report authored by Lorenzo Picotti on behalf of the Association International de Droit Pénal – XXI International Congress of Penal Law on "Artificial Intelligence and Criminal Justice", International Colloquium of Section I (Criminal Law-general part): "Traditional Criminal Law Categories and AI: Crisis or Palingenesis?", 2022, para 2. Hereinafter only "General Report".

By and large, there is a mismatch between the emergence of technology and regulation, and vice versa.[927] As it has been noted, "the legislative branch seems to be moving at a negligible speed compared to the technological advancements enforcing the perception that traditional regulation does not fit in this challenge".[928] Among possible causes, authors mention "the lack of a thorough and accurate definition of AI … which is aggravated by the fact that the definition changes as the technology evolves".[929] The debate on the regulation of AI presents two sides: on the one side, there are those who argue that regulation stifles innovation; on the other side, there are those who believe in "anticipatory policy-making".[930]

Furthermore, there is no general consensus on how liability standards should look like.[931] Moreover, there is a "huge gap between ethical guidelines and laws",[932] which lead to a surge in the development of principles related to "ethical AI" in the past 5 years.[933] Consequently, the slow progress of hard-law regulation does not come as a surprise, especially in field as delicate as criminal law.

---

927 E. Fosch-Villaronga & M. Heldeweg, "Regulation, I presume?" said the robot – Towards an iterative regulatory process for robot governance", *Computer Law & Security Review: The International Journal of Technology Law and Practice*, 2018, p. 2.

928 Patricia Gomes Rêgo de Almeida, Carlos Denner dos Santos, Josivania Silva Farias, Artificial Intelligence Regulation: a framework for governance, Ethics and Information Technology (2021) 23, 507.

929 Patricia Gomes Rêgo de Almeida, Carlos Denner dos Santos, Josivania Silva Farias, Artificial Intelligence Regulation: a framework for governance, Ethics and Information Technology (2021) 23, 508.

930 N. Boucher, European Parliamentary Research Service - Scientific Foresight Unit, "What if AI regulation promoted innovation?"; PE 729.515, 2022, p. 2.

931 A. Folberth et. al, Karlsruhe: Karlsruhe Institute of Technology (KIT), Institute for Technology Assessment and Systems Analysis (ITAS), "Tackling problems, harvesting benefits – A systematic review of the regulatory debate around AI", *KIT Scientific Working Papers*, Vol. 197, 2022, p. 11.

932 P. Gomes Rêgo de Almeida, C. Denner dos Santos & J. Silva Farias," Artificial Intelligence Regulation: a framework for governance", *Ethics and Information Technology*, Vol. 23, 2021, p. 508.

933 E.g., AI-HILEG*, Ethics Guidelines for Trustworthy Artificial Intelligence (AI),* 8 April 2019; OECD, *Recommendation of the Council on Artificial Intelligence,* OECD/LEGAL/0449. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; UNESCO, *Recommendation on the Ethics of Artificial Intelligence,* SHS/BIO/REC-AIETHICS/2021, https://unesdoc.unesco.org/ark:/48223/pf0000380455; Council of Europe - CEPEJ, *European Ethical Charter on the Use of AI in Judicial Systems Council of Europe*, 2018. Available at: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699. For a systematization of ethical guidelines and principles J. Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*", Berkman Klein Center for Internet & Society*, 2020 C. Rudschies, I. Schneider & J, Simon, "Value Pluralism in the AI Ethics Debate – Different Actors, Different Priorities", *The International Review of Information Ethics*, Vol. 29, 2021; A. Jobin, M. Ienca & E. Vayena, "The global landscape of AI ethics guidelines", N*ature Machine Intelligence*, Vol.1, 2019.

Nevertheless, this study identified five relevant inputs in the field of criminal regulation of AI systems:[934]

A) The Council of Europe's European Committee of Criminal Problems and the drafting of an " Instrument on Artificial Intelligence and Criminal Law";
B) The Singapore Penal Code Review Committee Report of 2018[935] and the Report on "Criminal Liability, Robotics and AI systems" drafted by the Singapore Law Commissions of 2021;[936]
C) The legislative reform of the French Road Act (*Ordonnance n° 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation*);
D) the "Automated Vehicles: joint report" drafted by the Law Commission of England and Wales and by the Scottish Law Commission;[937]
E) the amendment of the German Road Traffic Act.[938]

---

934 Part of these reflections were developed in Giannini & Kwik, 2023. More insights can be found in the AIDP's General Report for Section I. There, the author refers to an academic project which took place in Hungary and resulted into the creation of a draft model law addressing the issue which introduces the concept of "AI sanctionability" instead of criminal liability. The General Report also quotes also the Chinese National Report for Section I, where the following punishments for Strong AI systems are envisaged: "«Data Deletion» , namely deleting the data information on which the strong AI system relies to commit crimes, thus depriving it of the ability to commit previous crimes; «Program Modification», namely modifying the program of the strong AI system to restrict the strong AI system's learning ability and data acquiring ability, thus depriving it of its independent recognition and control ability and allowing it to commit acts only within the scope of human control; and «Program Deletion», namely removing all programs related to the strong AI system so that the intangible strong AI system, which depends on the program to survive, no longer exists". Moreover, the Chinese report focuses on robots and displays that "it is possible to apply penalty such as restriction of freedom, deprivation of freedom and destruction. The restriction of the space for them to conduct physical activity can restrict or deprive of their freedom. On basis of this, relevant ethical and legal norms can be re-introduced to them during the period when their freedom is restricted or deprived of, so as to educate and transform the strong artificial body. For tangible strong AI systems that cannot be educated and transformed, they can be physically destroyed". Thus, as the author of the General Report notes, GAI has not been developed (yet), hence the Chinese statements remain purely theoretical. See AIDP General Resolution Section I, 2022, pp. 8-9. See also B. Miskolczi & Z. "*Büntetőjogi kérdések az információk korában – Mesterséges intelligencia"*, *Big Data, profilozás. HVG-ORA*C, Budapest, 2018.

935 PCRC Report, 2018.

936 Singapore Academy of Law, 2021.

937 Law Commissions Report, 2022.

938 Bundestag, Gesetz zur Änderung des Straßenverkehrsgesetzes, 2021.

They present similarities and differences. To begin with, A is a cross-border tool; where instead B, C, D and E have a national scope of application. Moreover, B and D are – at the present moment – proposals to reform the law, where instead C and E have already been implemented. Additionally, C, D and E are sector-specific, as they regard exclusively the field of autonomous driving, while instead A and B have a broader scope of application.[939]

On a final note, a mention should be made to the EU, specifically to the European Proposal for a Regulation on Artificial Intelligence (Artificial Intelligence Act, "AIA") of April 2021. Notably, the AIA is an example of a "risk-based" approach to regulation: it divides AI uses according to different levels of risk. Each level of risk is accompanied by different obligations (which include quality management systems, CE conformity mark, and transparency duties). So far, there has not been any input by the EU regarding *criminal* liability rules in connection to AI systems, whereas in September 2022 the European Commission delivered a proposal for a Directive on adapting non contractual civil liability rules to artificial intelligence.[940]

7.2 GENERAL-SCOPE TOOLS

A) Council of Europe

In November of 2018, the Council of Europe ("CoE"), specifically the European Committee of Criminal Problems (CDPC),[941] organized a Thematic Session on AI and criminal law responsibility, whose focus was "the importance of a meaningful approach in legal systems across Europe to deal with the challenging questions posed by the increased presence of artificial intelligence in civil life" and it included, amongst its aims, the examination of "the scope and substance of an *international legal instrument to provide common standards for the criminal*

[939] For a more thorough comparison, see Giannini & Kwik, 2023.

[940] European Commission, *Proposal for a Directive of the European Parliament and of the Council On Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence* (*AI Liability Directive*), COM(2022), 496 final 2022/0303 (COD), 28 September 2022.

[941] The CDPC was established in 1958 by the Committee of Ministers under Article 17 of the Statute of the CoE. Its goal is to set European standards and principles via binding and non-binding legal texts in a variety of fields related to criminal law and procedure.

law aspects of automated technologies".[942] First and foremost, the CDPC highlighted, according to the *ultima ratio* principle, that the "*degree of harm*" and the "*importance of the obligation breached*" need to be taken into consideration before criminal liability is triggered.[943]

The CDPC's initial focus was the case of automated vehicles, notably on how "to set up rules governing any potential criminal liability in advance to ensure that in cases such as a car collision or a drone crash, no State will have to face an unclear legal situation due to unsuitable or out-of-date rules".[944] Nevertheless, it was decided to treat it as a "general-scope" tool since the CDPC's intention is to use AV merely as an example of AI deployments and the future legal instrument will presumably have a larger scope of application.[945]

Following the thematic session, the CDPC established the Working Group of Experts on Artificial Intelligence and Criminal Law ("Working Group") to assist the CoE in its research on criminal law and AI.[946] The Working Group, after a first meeting in March 2019, prepared a questionnaire to conduct a census of relevant national criminal norms applicable to automated vehicles (or other AI deployments).[947] The questionnaire starts off with an example case:

> Imagine that, for the first time, a vehicle equipped with an "autopilot system" can be used legally on highways in your country. The automated driving system must be used in harmony with the authorisation which requires – among other things – that the human driver is ready to take over the steering wheel within 20 seconds. To ensure the driver's fitness to take over, the producer installs a drowsiness detection system

---

[942] Council of Europe, European Committee on Crime Problems (CDPC), "Thematic session on artificial intelligence and criminal law the approach in Council of Europe member states the case of automated vehicles", Programme, CDPC(2018)18, 28 November 2019, p. 2. Available at: https://rm.coe.int/cdpc-2018-18-draft-programme-thematic-session-artificial-intelligence-/16808e64ab.

[943] Council of Europe, CDPC, Concept Paper, 2018, pp. 4-5.

[944] Ivi, pp. 4-5.

[945] Council of Europe, CDPC, Working Group of Experts on Artificial Intelligence and Criminal Law, "Questionnaire concerning Artificial Intelligence and Criminal Justice (using the example of Automated Driving)", CDPC(2019)8FIN (2019), 19 May 2019, p. 3. Available at: https://rm.coe.int/cdpc-2019-8fin-questionnaire-artificial-intelligence-and-criminal-just/168094c8fa.

[946] Council of Europe, CDPC, Working Group of Experts on Artificial Intelligence and Criminal Law, "Working paper II", CDPC(2019)7, 27 March 2019, p. 3. Available at: https://rm.coe.int/cdpc-2019-7-working-paper-ii-for-cdpc-expert-group-meeting-on-artifici/16809372a5.

[947] Council of Europe, CDPC, Working Group of Experts, "Questionnaire concerning Artificial Intelligence and Criminal Justice", 2019, p. 3.

> monitoring the driver (seating position, face and especially eye movements) and stores the data with a cloud service provider. During the first months of operation of such cars, it turns out that a certain weather phenomenon in your country (be it morning mist, a sandstorm, midday sun or garbage thrown on the roadside) triggers faulty reactions in the driving assistant's system – especially false-braking, i.e. braking for the wrong reason, for instance a plastic bag drifting in the wind. The producer and all component suppliers do their very best to fix the problems. However, it is clear to everyone involved that the cars will need time to adjust to particular local conditions.[948]

Let us imagine that the car hits a human being while driving on autopilot, causing her death, and that it is further established that not only the car's sensors were defective, but also that the braking assistant had a severe software defect. Nevertheless, it is not possible to ascertain which defect caused the accident. In this case, the questionnaire asks whether the domestic law of the member country knows the concept of "contributory negligence" and, in case of a positive answer, whether it is seen as a matter of theories of causation or of complicity (or collaboration in negligence).[949]

The questionnaire continues by investigating the concept of societal risk. Let us suppose that "it could be proven that the car's sensors did not pick up the victim, most likely because he/she held a bag at arm's length and the engineers had 'tuned out' bag images from the sensors' vision in order to prevent 'false braking'".[950] In a case like this, "criminal justice systems may provide an option to forgo criminal prosecution, arguing that in the light of the overall social benefits a particular type of risk taking should not be punished even if harm is caused as long as the person in question does its best to comply with all requirements of safety and security".[951] As a matter of fact, this is the dominant school of thought with regards to airbags in cars ("although there is a minimal risk that this safety device might open because of a pothole and kill a passenger, it will moreover save lives in many situations").[952]

---

[948] Ivi, p. 3.
[949] Council of Europe, CDPC, Working Group of Experts, "Questionnaire concerning Artificial Intelligence and Criminal Justice", 2019, p. 5.
[950] Ibid.
[951] Council of Europe, CDPC, Working Group of Experts, "Questionnaire concerning Artificial Intelligence and Criminal Justice", 2019, p. 5.
[952] Ibid.

The assessment of the answers of the questionnaire reported a tendency of states to remain rooted in traditional notions and liability schemes, even when adopting new regulations. Notably, none of the 36 member states that filled the questionnaire[953] "opted for the creation of a new legal notion (such as an "e-personhood").[954] Thus, the Working Group concluded that the potential capacity of AI systems "to challenge the community's trust in the validity of the law"[955] could warrant "more functional approach in criminal justice, and possibly an evaluation of the option of a e-personhood".[956] Moreover, the Working Group advised for the introduction of new regulation which should update or create new concepts of liability, including corporate liability (but not necessarily *criminal* liability).[957]

Finally, in September 2020 the Working Group published a feasibility study on a future Council of Europe instrument on AI and criminal law.[958] The purpose of the feasibility study was to address whether an ad hoc Council of Europe committee of experts should be set up to prepare a draft instrument setting common criminal law standards on different relevant issues raised by vehicles driving autonomously (or other AI deployment).[959] The answer to said question is positive: according to the Working Group, even though issues of criminal liability are a matter of national jurisdictions, in the case of AI they must be regulated "within an international and collaborative framework".[960]According to the CPDC, the reasons for

---

953 These are: Andorra, Armenia, Austria, Azerbaijan, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Italy, Moldova, Monaco, Montenegro, North Macedonia, Norway, Latvia, Lithuania, Luxembourg, Poland, Portugal, Romania, Russia Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey and Ukraine.

954 Council of Europe, European Committee on Crime Problems (CDPC), 'Assessment of the answers to the questionnaire on artificial intelligence and criminal justice (using the example of Automated Driving)', CDPC(2019)17 (2019), 5. Available at: https://rm.coe.int/cdpc-2019-17-draft-assessment-of-the-answers-to-the-questionnaire-on-a/168098e24c.

955 Council of Europe, CDPC, "Assessment of the answers to the questionnaire on artificial intelligence and criminal justice (using the example of Automated Driving)", CDPC(2019)17 (2019), 7 November 2019, p. 12. Available at: https://rm.coe.int/cdpc-2019-17-draft-assessment-of-the-answers-to-the-questionnaire-on-a/168098e24c.

956 Ibid.

957 Council of Europe, CDPC, "Assessment of the answers to the questionnaire", 2019, p. 12.

958 Council of Europe, CDPC, Working Group on AI and Criminal Law & CDPC Secretariat, "Feasibility study on a future Council of Europe instrument on artificial intelligence and criminal law", CDPC(2020)3Rev (2020), 4 September 2020. Available at: https://rm.coe.int/cdpc-2020-3-feasibility-study-of-a-future-instrument-on-ai-and-crimina/16809f9b60.

959 Council of Europe, CDPC, Working Group on AI and Criminal Law & CDPC Secretariat, "Feasibility study", 2020, p. 4.

960 Ivi, p. 9.

creating an international legal instrument are manifold: it would foster the development of consistent legislation across Europe ("[it is not a question of devising a whole new system of liability that would overturn the criminal law of each member state, but rather of agreeing on a general framework for criminal law and AI deployment within which state-wide regulations could be developed");[961] it would bring legal certainty to European citizens and assist free movements across national borders; it would facilitate co-operation between different states on AI-related matters, facilitating the exchange of data. Based on the results of the questionnaire, the Working Group decided to exclude the issue of creating a legal personality for AI systems which would be relevant in criminal matters.[962]

Conclusively, the CDPC created an ad hoc committee tasked with the drafting of such instrument (Drafting Committee to elaborate an instrument on Artificial Intelligence and Criminal Law, "CDPC-AICL"). The potential nature of the instrument which shall be adopted was discussed during the first two meetings of the CDPC and the members agreed for the drafting of a Recommendation.[963] The deadline to submit a draft of the legal instrument to the Committee of Ministers has now been set for December 2023.[964]

B) Singapore

The proposals analyzed in this paragraph are a paramount example of the Singaporean[965] strive to become key normative players in the field of AI. Indeed, Singapore

961 Council of Europe, CDPC, Working Group on AI and Criminal Law & CDPC Secretariat, "Feasibility study", 2020, p. 10.
962 Ivi, p. 12.
963 Council of Europe, CDPC, "1st meeting of the Drafting Committee to elaborate an instrument on Artificial Intelligence and Criminal Law (CDPC-AICL)", CDPC-AICL(2021), 17 November 2021, p. 3. Available at: https://rm.coe.int/cdpc-aicl-2021-1-1st-meeting-report-15-16-nov-2021/1680a49c99.
964 CDPC, "Terms of Reference", Extract from CM(2021)131-addrev, p. 2. Available at: https://rm.coe.int/cdpc-en-terms-of-reference-cm-2021-131-addrev/1680a4b41a; Council of Europe, CDPC, "2nd meeting of the Drafting Committee to elaborate an instrument on Artificial Intelligence and Criminal Law (CDPC-AICL)"; CDPC-AICL(2022)2, 9 June 2022, p. 3. Available at <https://rm.coe.int/cdpc-aicl-2022-2-2nd-meeting-report/1680a6e1ff>.
965 The legal system in Singapore is based on British common law. Its criminal justice system is adversarial and pluralistic since it incorporates elements of Malay customary law and Muslim law. Its Penal Code and Criminal Procedure Code have been heavily influenced by Indian legal culture. Cfr. M. Nalla "Singapore", *World Factbook of Criminal Justice Systems*, 1993. Available at: https://bjs.ojp.gov/content/pub/pdf/wfbcjss.pdf; S. Yeo, N. Morgan & C. Wing Cheong, *Criminal Law in Malaysia and Singapore,* 2nd Ed., LexisNexis, 2012.

is seeking to establish itself as an AI "rule of law hub",[966] by means of introducing regulation "to attract and encourage AI innovation".[967] This effort is substantiated into two different proposals: the Singapore Penal Code Review Committee (PCRC) Report of 2018[968] (and, more specifically, the proposal to introduce two new offences relating to computer programs);[969] and the Singapore Academy of Law's[970] Law Commission Report on Criminal Liability, Robotics and AI Systems of 2021 ("the SAL report").

Let us start from the first. The PCRC Report, amongst other things, suggests the introductions of two new offences. According to the first offense,

> (1) Whoever makes, *alters* or *uses* a computer program so rashly[971] or negligently as to endanger human life, or to be likely to cause hurt or injury to any other person, or knowingly or negligently *omits* to take such order with any computer program under his care as is sufficient *to guard against any probable danger* to human life from such computer program, shall be punished with imprisonment for a term which may extend to one year, or with fine which may extend to $5,000, or with both.

---

966 Chesterman, 2021, p 5.

967 Ivi, p. 5. In 2018, the Singapore Penal Code Review Committee (acknowledged that "[b]eing the global first-mover" on rules regarding criminal liability related to AI systems might "impair Singapore's ability to attract top industry players in the field of AI". Nevertheless, it further advised the Singaporean government to "actively explore and develop a suitable framework to address the issue of criminal liability for harm caused by computer programs. This should be done in the broader context of Singapore's developing regulatory framework for AI". PCRC Report, 2018, p. 29.

968 The PCRC was established by the Singaporean of Home Affairs and Ministry of Law in 2016 to review the Singapore Penal Code and make recommendations on how to reform it. It completed its review in 2018.

969 The PCRC includes AI in the term "computer programs". See PCRC Report, 2018, p. 27, Para. 2.

970 The Singapore Academy of Law ("SAL") is a private organization that was established in 1988 with the goal of making Singapore the "legal hub of Asia". It is led by a Senate which is headed by the Chief Justice and comprises of the Attorney-General and the Supreme Court Bench. See: https://www.sal.org.sg.

971 In the Singapore criminal legal system, rashness constitutes a form of culpability, akin to negligence, that results from a failure to exercise a reasonable level of care and caution in one's actions. Rash acts are characterized by imprudent or impulsive behavior, without taking appropriate safety measures, whereas negligence typically arises from routine actions that are commonly understood to pose some degree of danger. S. JLS, "The Continuing Confusion Over Section 304A of the Singapore Penal Code", *Singapore Journal of Legal Studies,* 2015, p. 144.

(2) For the purposes of this section, a person uses a computer program if he causes a computer holding the computer program to perform any function that —

(a) causes the computer program to be executed; or
(b) is itself a function of the computer program.[972]

(3) For the purposes of this section, a computer program is under a person's care if he has the lawful authority to use it, cease or prevent its use, or direct the manner in which it is used or the purpose for which it is used.

The proposed offense would target two groups of individuals: those who create and modify computer programs, and those who use them.[973] It establishes a crime of endangerment:[974] the new offense would punish a conduct of "risk-creation"[975] regardless of the verification of harm. The realization of harm, whether resulting in physical injury or death, would trigger the application of other offenses of the Singapore Penal Code (such as articles 304A or 337). As noticed by the PCRC itself, such an offense does not include scenarios in which, on the one hand, the harm caused does not result in physical harm or death and, on the other, the "user" is not aware that the (specific) harm will occur, "either because the program is capable of learning new behaviors on its own or because the program is designed to act random".[976]

To fill this apparent *lacuna* the PCRC proposes the adoption of a second offense:

---

[972] PCRC Report, 2018, p. 30.
[973] PCRC Report, 2018, p. 30, para. 13.
[974] Crimes of endangerment are criminal offenses that punish acts or omissions that create a significant risk of harm to others, regardless of whether that risk is ultimately realized. They are characterized by a failure to show proper concern for the safety of others. While the actus reus elements of endangerment offenses may not present significant challenges, there is ongoing debate regarding the appropriate mens rea connection, specifically whether it should be viewed as a form of strict liability or fault-based culpability. From a perspective that aligns with principles of culpability, the offender should be held liable for their indifference to the risk they created, specifically for displaying an attitude of disregard for legally protected interests. See A. Duff, & T. Hörnle, "Crimes of Endangerment", in Ambos K. et al. (Eds.), *Core Concepts in Criminal Law and Criminal Justice,* Vol. 2, Cambridge University Press, 2022, pp. 132-166; R. A. Duff, "Criminalizing Endangerment", *La. L. Rev.*, Vol. 65, 2005, pp. 944-945.
[975] PCRC Report, 2018, p. 13, para. 30.
[976] PCRC Report, 2018, p. 30, para. 14.

> (1) Where a computer program —
> (a) produces any output, or
> (b) performs any function,
>
> that is likely to cause any hurt or injury to any other person, or any danger or annoyance to the public, and the computer program *is under a person's care*, if that person knowingly *omits to take reasonable steps to prevent* such hurt, injury, danger or annoyance, he shall be punished with imprisonment for a term which may extend to one year, or with fine which may extend to $5,000, or with both.[977]

This formulation would attach responsibility also to those overseeing an AI system who were not aware of the existence of any kind of risk connected to it. Indeed, differently from the first offense, the duty of care ("to take reasonable steps to prevent such hurt, injury, danger or annoyance") is not tied to the knowledge of a general risk ("that the computer program is likely to cause any hurt or injury to any other person, or any danger or annoyance to the public"). The way the offense is formulated, specifically the vagueness of the concept of being "under a person's care", seems to entail "that the general risk of harm could also be an objective and intrinsic characteristic of the computer program, ie, one that is *independent from any subjective evaluation of the culpable agent*. […]" and, therefore, that a user could be liable for not acting to mitigate the risks connected to an innate feature of AI systems.[978]

Let us move now to the second proposal, that is, the SAL Report of February 2021. The SAL report looks into the potential dangers that autonomous Robotic and Artificial Intelligence systems (RAI) can pose to people and property. It specifically focuses on instances where harm occurs and whether or not Singaporean criminal laws should apply and how criminal liability should be determined. The report acknowledges that there are various uses for these systems, each with its own unique risks and benefits, making it difficult to have a one-size-fits-all approach to criminal liability. [979] Therefore, the report examines two factors: first, whether or not there is a human "involved in operating, affecting, or overseeing

[977] PCRC Report, 2018, pp. 31-32.
[978] Giannini & Kwik, 2023, pp. 22-23.
[979] SAL Report, 2021, p. 10, para. 14.

the RAI system); second, "where such a human is involved, whether they intended or knew the harm would occur".[980]

The Committee argues that the first issue would be to identify the "user-in-charge".[981] In cases where the level of automation is lower than level of human oversight exercised (i.e., "partial automation"), the user-in-charge would be the subject who "directly controls or is responsible for determining the actions of the RAI systems".[982] In cases of highly (yet not fully) automated RAI systems, the user-in-charge would be either the subject who is ultimately responsible for deciding/approving a particular action, the one who maintains control over the system's decision-making process, or the one who is specifically obligated to intervene to control the system's action in a given scenario.[983] Considering this, the SAL Report then distinguishes between two different situations: first, cases of intentional criminal use of, or interference with, the RAI system; second, cases of *non-intentional* harms. Let us focus on the latter, as the use of AI systems as (very sophisticated) tools to commit crimes is outside the scope of this research.

As described in the SAL Report, in order to establish negligence the Singapore Penal Code[984] sets two conditions: (a) determining what an objective "reasonable person" would do in a given circumstance, and (b) proving that that standard was breached in the specific case.[985] When addressing harm caused by RAI that falls within the realm of existing negligence-based offenses, it would be up to courts to "apply or adapt existing criminal negligence standards, or – in the absent of precedent – define new one".[986] Furthermore, a new negligence-based offense might be created to cover all negligent actions that result in harm from RAI systems, setting the reasonable conduct threshold. As a result, the risk of such a broadly applicable rule is that it may not be sufficient to capture RAI system actions

---

980 SAL Report, 2021, pp. 10-11, para. 1.5.

981 SAL Report, 2021, p. 23, para. 4.1. "User-in-charge" is the same term adopted by the UK Law Commissions, as it will be outlined in the next paragraph. The Committee addresses this overlap and states that "[w]hile utilising the same term, the definition of 'user-in-charge' adopted here differs from that utilised by the UK Commissions in the specific context of automated vehicles (although its 'users-in-charge' would equally fall within the definition utilised here)". SAL Report, 2021, p. 23, No. 34.

982 Ivi, p. 23, para. 4.3.

983 SAL Report, 2021, pp. 23-24, para. 4.3.

984 "Whoever omits to do an act which a reasonable person would do, or does any act which a reasonable person would not do, is said to do so negligently". Singapore Penal Code, 1871, S 26F(1) PC.

985 SAL Report, 2021, p. 30, para. 4.24.

986 SAL Report, 2021, p. 31, para 4.26.

that have never occurred before, i.e., conducts for which "existing precedents are inappropriate or for which there is no existing precedent at all".[987]

Additionally, the SAL Report considers legislating the adoption of industry- or technology-specific norms of conduct. One of the examples mentioned by the drafters is the one of AVs: the SAL Report argues that regulation might define specific situations in which the user-in-charge must assume control of the vehicle, such as when a route is temporarily blocked due to a traffic accident.[988] In this regard, the SAL Report's strategy is distinct from that of the UK Law Commissions. As it will be shown, the latter concentrates on the suitability of the single Autonomous Driving System feature of the vehicle, to be validated through an authorization procedure, rather than on external circumstances (e.g., an accident) and their impact on the obligation of the operator to intervene. In fact, according to the UK model the user-in-charge is not responsible for *any* "dynamic driving offence"[989] or civil sanction that occurs while such a feature is activated.

It is worth noting that the SAL Commission highlights the importance of every stage of the AI deployment process (data preparation, training of the mode, choosing the relevant model[s], the environment where the RAI system is deployed) as probable causes of the realization of a criminal offense.[990] The SAL Report also highlights the fact that harm could result from the amount, accuracy, and quality of the training data as well as the RAI system's architecture (i.e., its code). The importance of comparing the environments in which the system was trained and deployed, on the one hand, and the real-world data that the RAI system had collected at the time the damage was committed, on the other, should also be considered.[991] By doing so, the SAL Commission devotes attention to factors that are typically overlooked by scholarly discourse, as stated in Ch. 3.5.

The SAL Commission, reasoning under the premise that criminal negligence may not (always) be the solution, recommends four fixes for what was referred to as "negligence failures".[992] The first one is the creation of a new form of legal personality for RAI systems,

---

987 SAL Report, 2021, p. 31, para. 4.27

988 Ibid.

989 See below at C) for a definition of dynamic offences.

990 SAL Report, 2021, p. 32, para 4.32.

991 Ibid.

992 "Negligence failures can be defined as situations in which the classical building blocks of negligence, ie, risk taking, foreseeability, and awareness, struggle to identify a liable human being to whom we can attribute AI-caused harm". Giannini & Kwik, 2023, p. 3.

so that criminal liability could be imposed *directly* on the RAI system itself.[993] Doing so would diminish "the extent to which it is necessary to get 'under the bonnet' of an RAI system, and identify which specific part or parts of that system caused its decision to act as it did, and which of the (potentially numerous) parties involved in the system's development and deployment should be held responsible there for".[994] It follows that holding AI systems directly liable would have an "indirect penalizing effect on those responsible for or profiting from the RAI system, while minimizing the need to prove that the harm was attributable to specific natural persons or corporations".[995] The SAL Committee ultimately discards this option since it considers the arguments against separate personality for RAI systems more compelling:

> We consider that criminal laws should continue to be formulated on the basis that such laws are intended to shape or impact human behaviour. It is not fully clear, for example, how imposing criminal liability (and a sanction) on an RAI system directly would deter the system itself from causing harm. And to the extent that the objective would be to deter or penalize those responsible for the RAI system, rather than the system itself, we also take the view that such ends could equally be achieved through alternative mechanisms that do not require the creation of wholly new forms of legal personality (with all the disruption to the existing legal framework that that would necessarily entail).[996]

In contrast, the second and third options refer to the offenses from the PCRC Report. The Committee observes that, despite the fact that the offenses may indeed address negligent failures, they do not sufficiently define the boundaries of the duty of care related to the AI system. Specifically, more work needs to be placed into defining precisely what "a rash or negligent act" or "failure to take reasonable steps in any given case" means.[997]

---

[993] SAL Report, 2021, p. 36, para. 4.41.
[994] SAL Report, 2021, p. 37, para. 4.43.
[995] SAL Report, 2021, p. 37, para. 4.44.
[996] SAL Report, 2021, p. 38, para. 4.47.
[997] SAL Report, 2021, p. 40, para. 4.54.

The fourth alternative is to use workplace safety legislation as a model. This represent an original suggestion, which resembles Diamantis' "Employed Algorithms" doctrine.[998] Notably, workplace safety is one of the subjects where criminal law identifies certain "centers of imputation" of liability who are entrusted with the protection of legal goods and, as a consequence, they are deemed responsible if they fail to do so – or do so poorly.[999]

Imputation mechanisms related to risk governance, especially when they involve complex structures such as corporations or the AI-development process, are a slippery slope. The more distance there is from the specific (natural) cause of a harm, the more issues with regards to the principle of culpability arise. This type of liability is usually deeply rooted into statutory duties "to take all reasonably practicable measures to avoid the harm (including, for example, the adequacy of the protective processes and systems the entity had in place)"[1000] and less focused on the determination of "the specific cause of the harm"[1001], or on the negligence of a (natural or legal) subject.[1002] In other words, "the prosecution need not to prove a direct or scientifically precise causation between the harm caused by the RAI and a particular breach of duty".[1003]

Now, the SAL Report suggests the introduction of a system similar to the one provided by the Workplace Safety and Health Act ("where a duty is imposed on specified entities to take, so far as is reasonably practicable, such measures as are necessary to avoid harm").[1004] Accordingly, one could think of imposing a duty on the entity which is "best placed – on the bases of their 'proximity' to the RAI system and its operation, and their resources – to take action (i.e., to prevent, address and rectify dangers posed by RAI systems) and to change future outcomes".[1005] As it will be explained in a moment, the concept of "proximate liable entity" recurs also in the UK Law Commissions Report, specifically with reference to the Automated Driving System Entity (ASDE).[1006]

---

998 See above at Para. 6.4.2.

999 Cfr., *ex multis*, E. Scaroina, "La responsabilità penale del datore di lavoro nelle organizzazioni complesse", *Sistema penale*, 2021.

1000 SAL Report, 2021, p. 42, para. 4.62.

1001 Ibid.

1002 SAL Report, 2021, p. 42, para. 4.62.

1003 Ibid.

1004 SAL Report, 2021, p. 41, para. 4.58.

1005 SAL Report, 2021, p. 41, para. 4.59.

1006 See below at C).

While the SAL Commission acknowledges the issues which such a model may pose, it concludes by stating that its introduction might be justified by the seriousness of the risks posed by the RAI and by the citizens' need of an accountability framework. Particularly, introducing RAI-related duties in the field of workplace safety ("where distinct statutory duties are already imposed and the appropriate policy balance has thus already been considered and determined")[1007] would not entail reinventing the wheel, rather, it would mean to "review (and as necessary amend) those existing laws to ensure that occupiers and employers may equally be held responsible for harm resulting from the autonomous operation of RAI systems in their workplaces".[1008] Finally, it would amount to "a policy judgment for lawmakers, balancing, in particular, demands for accountability with the desire not to unduly stifle innovation and impede the societally-beneficial development and use of RAI systems".[1009]

7.3 SELF-DRIVING TRAGEDIES: AV-SPECIFIC TOOLS

The realm of driving automation has been the stage of numerous self-driving tragedies[1010] in recent years, which have called into play the application of criminal law. It is worth mentioning a few of them to provide context for the three specimens of regulation that this paragraph is about to examine.

In May 2016, a collision between a Tesla Model S 70D car,[1011] which was operated using Traffic-Aware Cruise Control and Autosteer lane-keeping systems (SAE Level 2),[1012]

1007 SAL Report, 2021, p. 43, note 84.

1008 Ibid.

1009 SAL Report, 2021, p. 42, para. 4.63.

1010 The term is taken from Smiley, 2022.

1011 NTSB, Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida, May 7, 2016, Highway Accident Report NTSB/HAR-17/02, 2017, p. 9 ("NTSB Report Florida"). Available at: https://www.ntsb.gov/investigations/accidentreports/reports/har1702.pdf.

1012 "TACC is an adaptive cruise control system that maintains the set cruise speed, applies brakes to preserve a predetermined following distance when approaching a slower-moving vehicle ahead of the Tesla, and accelerates to the set cruising speed when the area in front of the Tesla is no longer obstructed. Autosteer automatically steers the car to keep it within its lane of travel. In short, TACC provides longitudinal control (acceleration and deceleration) and Autosteer provides lateral control (steering) of the car within the lane, making the Tesla Autopilot consistent with an SAE International (SAE) Level 2 automated vehicle system". Ibid.

and a truck-tractor, caused the death of the occupant of the car nearby Williston, Arizona. The investigation on the crash was directed by the National Transportation Safety Board (NTSB), after it learnt of a defect investigation related to automatic emergency braking Autopilot systems of the Tesla Models S and X of the National Highway Traffic Safety Administration (NHTSA).[1013] The NTSB is an independent federal agency established in 1967 and is dedicated to promoting aviation, railroad, highway, marine, and pipeline safety. Its mission does not include the assignment of fault or blame for an accident or incident.[1014] Rather, "accident/incident investigations are fact-finding proceedings with no formal issues and no adverse parties … and are not conducted for the purpose of determining the rights or liabilities of any person".[1015] As such, it represents an instance of an entity conducting *non-punitive* and *non-blameworthy* investigations as happens in the field of aviation.[1016] In its investigation on the Florida crash, the NTSB focused on the Autopilot system and concluded that the probable cause of the crash was "the truck driver's failure to yield the right of way to the car, combined with the car driver's *inattention due to overreliance on vehicle automation*, which resulted in the car driver's lack of reaction to the presence of the truck".[1017] Moreover, "[c]ontributing to the car driver's overreliance on the vehicle automation was its *operational design*, which *permitted his prolonged disengagement from the driving task* and his use of the automation in ways inconsistent with guidance and warnings from the manufacturer".[1018] Notwithstanding the fact that Tesla's automated vehicle control system "was not designed to, and did not, identify the truck crossing the car's path or recognize the impending crash"[1019] and, consequently, it "did not reduce the car's velocity, the forward collision warning system did not provide an alert, and the automatic emergency braking did not activate";[1020] that it was not designed as to automatically restrict its operation in conditions for which they are not designed (such as driving on a highway); and that, overall, Tesla failed

[1013] National Highway Traffic Safety Administration (NHTSA), Office of Defects Investigation, Investigation PE 16-007, 2017. Available at: https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF.
[1014] NTSB Report Florida, 2017, p. 3.
[1015] 49 C.F.R. § 831.4.
[1016] See above Ch. 4.1.
[1017] NTSB Report Florida, 2017, vi.
[1018] Ibid.
[1019] NTSB Report Florida, 2017, p. 41.
[1020] Ibid,

to consider adequately the human element;[1021] there were no investigations on Tesla's (criminal) involvement in the accident. Notably, it was reported that the Dutch vehicle authority (*RijksDienst voor het Wegverkeer* or "RDW") requested information on the crash from the NHTSA, as it regarded autopilot features that had been approved by the RDW for its use in Europe.[1022] Moreover, in July 2020, a Munich regional court (*Landgericht München I*) banned Tesla from using the terms "autopilot" and "full potential for autonomous driving" in the sale of his Tesla model 3 cars, since it mislead consumers into believing that these vehicles were technically capable, and legally authorized, to drive without any human intervention (i.e., level 5 automation), when in reality they only reach level 2 automation.[1023]

On March 18, 2018, Elaine Herzberg was struck and killed by an automated Uber test car (SAE level 3) carrying a human driver in Tempe, Arizona.[1024] According to NTSB's Report,[1025] since the ADS did not recognize the pedestrian as a jaywalker, it detected her

---

[1021] Tesla failed to recognize that "monitoring steering wheel torque provides a poor surrogate means of determining the automated vehicle driver's degree of engagement with the driving task"; moreover, the way its Autopilot System "monitored and responded to the driver's interaction with the steering wheel was not an effective method of ensuring driver engagement"; finally, it failed to consider the risk of driver overreliance on its system and the possibility of a lack of understanding of system limitations. NTSB Report Florida, 2017, p. 45.

[1022] Reuters, "Dutch vehicle authority seeks answers on fatal Tesla crash", 14 July 2016. Available at: https://www.reuters.com/article/tesla-authority-dutch-idINL8N1A03KF.

[1023] "(a) *Durch die Verwendung des Wortes "Autopilot" suggeriert die Beklagte aus Sicht der angesprochenen Verkehrskreise, die von ihr vertriebenen Fahrzeuge seien in der Lage, vollständig autonom zu fahren. Jedenfalls besteht eine hinreichende Gefahr, dass die angesprochenen Verkehrskreise den Begriff dahingehend missverstehen.*
(b) Die Formulierung "Volles Potenzial für autonomes Fahren" erweckt bei den angesprochenen Verkehrskreisen die Vorstellung, das von der Beklagten vertriebene Fahrzeug sei technisch in der Lage, vollkommen selbständig zu fahren oder zumindest die Herstellung dieser Eigenschaft sei ohne Weiteres durch geringfügige Modifikationen (Upgrades) erreichbar. Die Verwendung des Wortes "Potenzial" werden die angesprochenen Verkehrskreise entweder dahingehend verstehen, dass bei dem Fahrzeug eine technische Grundausstattung vorhanden ist, die ohne erhebliche Zwischenschritte und Investitionen ein Erreichen von Level 5 der in den USA und Europa branchenüblichen Klassifikation zum autonomen Fahren möglich macht. Zudem besteht die hinreichende Gefahr, dass die angesprochenen Verkehrskreise die in Frage stehende Formulierung dahingehend verstehen, der Nutzung der beworbenen Funktionen stünden allein regulatorische Hürden entgegen". LG München I, Endurteil vom 14.07.2020 - 33 O 14041/19, paras. 97-98.

[1024] Uber was conducting a preliminary test of a "Level 4" vehicle automation level as described by the National Highway Traffic Safety Administration. Cfr. A. DeArman, , "The Wild, Wild West: A Case Study of Self Driving Vehicle Testing in Arizona", *Arizona Law Review*, Vol. 61, 2019, p. 988, Note 28. Level 4 is "high automation" and entails that the system is fully responsible for driving tasks within limited service areas, while occupants act only as passengers and do not need to be engaged. When the system is engaged, it handles all the driving tasks and the driver is not required to maneuver the vehicle. Source: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety.

[1025] NTSB Report Tempe, 2018.

only about 6 seconds before the collision. Notably, the system was designed with little regard for jaywalking pedestrians. The human-in-the-vehicle was assigned the duties of "overseeing the system's operation, monitoring the driving environment, and, if necessary, taking control of the vehicle and intervening in an emergency".[1026] In other words, the vehicle operator was contemplated as the "primary countermeasure in an emergency situation"[1027] as she was expected to "recognize the hazard, to take control of the vehicle and to intervene appropriately".[1028] The report mentions the probable cause of the crash was the failure of the vehicle operator to monitor the driving environment and the operation of the ADS, since she was looking at her mobile phone. Contributing to the crash were Uber Advanced Technologies Group's (1) inadequate safety risk assessment procedures, (2) ineffective oversight of vehicle operators, and (3) lack of adequate mechanisms for addressing operators' automation complacency—all a consequence of its *inadequate safety culture*.[1029] Moreover, the same report recognizes that "[c]onsidering the roadway geometry, the sight distance, and the lighting in the crash area, the vehicle operator, had she been attentive, would have had *2 to 4 seconds* to detect and initiate a response to the crossing pedestrian to prevent the crash."[1030] Hence, an attentive behavior would have "likely"[1031] avoided the crash or mitigated its impact. Yet, the NTSB also acknowledged a crucial aspect, i.e., the fact that "[w]hen it comes to the human capacity to monitor an automation system for its failures, research findings are consistent—humans are very *poor* at this task"[1032] and that Uber failed to address these risks adequately. Specifically, the NTSB concluded that it was the vehicle operator's prolonged visual distraction, "a typical effect of automation complacency",[1033] that caused her to disengage with the vehicle and therefore lead to her failure to detect the pedestrian in time to avoid the collision. In a letter of 4 March 2019, the Yavapai County Attorney stated that there was no basis for criminal liability for the Uber corporation arising from the matter,[1034]

[1026] Ivi, p. 8.
[1027] NTSB Report Tempe, 2018, p. 14.
[1028] Ibid.
[1029] NTSB Report Tempe, 2018, pp. 57-59.
[1030] Ivi, p. 43.
[1031] NTSB Report Tempe, 2018, p. 57.
[1032] NTSB Report Tempe, 2018, p. 44 [emphasis added].
[1033] Ibid.
[1034] Yavapai County Attorney, "Re: Rafael Vasquez / Uber Corporation, Tempe Police Department #2018-32694", 4 March 2019. Available at: https://s3.documentcloud.org/documents/5759641/UberCrashYavapaiRuling03052019.pdf.

while the back-up driver in the Uber car was charged and indicted with a count of negligent vehicular homicide[1035] by a Maricopa County Grand Jury on 27 August 2020. [1036]

In a criminal trial which was supposed to commence in Los Angeles on November 15, 2022, Kevin George Aziz Riad, a Tesla Model S driver, is facing two charges of vehicular manslaughter with gross negligence.[1037] Specifically, Riad was behind the wheel when its car ran a red light while exiting a freeway and crashed into a Honda Civic, killing its two passengers. Allegedly, Tesla's Autopilot was engaged at the time of the crash.

It has also been reported that Tesla Inc. is now the subject of a criminal investigation by the U.S. Department of Justice ("DOJ").[1038] Presumably, the DOJ is investigating Tesla's marketing claims regarding the capabilities of its Autopilot, which could have led drivers to "imbue customers with a false sense of security, inducing them to treat Teslas as truly driverless cars and become complacent behind the wheel with potentially deadly consequences".[1039]

These three real life scenarios work as a useful display of the complex relationship which can occur between a human behind the wheel, an automated vehicle, and a lethal accident. The three examples which will be analyzed now are an attempt at simplifying this very tangled web.

C) France

In 2021, the French government adopted an *ordonnance* which amended the French Road Code by introducing specific provision on the criminal liability applicable to the circulation of a vehicle with driving delegation. We can summarize its most relevant passages as follows.

[1035] See A.R.S. §§ 13-1101, 13-1102, 28-3001, 28-3004, 28-3005, 28-3315, 13-701, 13-702, and 13-801.

[1036] State of Arizona, "Indictment 785 GJ 251", 27 August 2020. Available at: http://www.maricopacountyattorney.org/DocumentCenter/View/1724/Rafael-Vasquez-GJ-Indictment.

[1037] T. Krisher & S. Dazio, "Felony Charges Are 1st in a Fatal Crash Involving Autopilot" *AP NEWS,* Los Angeles, 18 January 2022. Available at: https://apnews.com/article/tesla-autopilot-fatal-crash-charges-91b4a0341e07244f3f03051b5c2462ae.

[1038] M. Spector & D. Levine, "Exclusive: Tesla faces U.S. criminal probe over self-driving claims", *Reuters,* 27 October 2022. Available at: www.reuters.com/legal/exclusive-tesla-faces-us-criminal-probe-over-self-driving-claims-sources-2022-10-26/.

[1039] Ibid.

First, the *ordonnance* added article L. 123-1 to the French Road Code, which stipulates that art. L. 121-1 of the French Road Code does not apply to drivers who commit infractions resulting from the manoeuvre of a vehicle whose driving functions have been delegated to an automated driving system, provided that the system was in dynamic control of the vehicle at the time of the offence.[1040] Apropos, art. L. 121-1 provides that "The driver of a vehicle shall be criminally liable for violations committed while operating said vehicle". Consequently, art. L. 123-1 introduces an immunity clause which shields drivers from criminal liability whilst certain functions of driving (i.e., "dynamic driving", which is not expressly defined in the Road Act), are performed by an AI system.[1041]

The second paragraph of art. L.123-1 proscribes that the driver must be *constantly* in a state and in a position to respond to a request to take control of the automated driving system. Indeed, according to the new art. L. 319-3, the AI system must alert the human driver whenever it is capable of exercising dynamic control of the vehicle in accordance with its conditions of use (para I). Moreover, after the driver has taken the decision to activate the automated driving system after the alert, the AI system must also perform a so-called *demand de reprise* anytime it perceives that it will not be able to perform successfully (para II). [1042]

[1040] "*Art. L. 123-1. Les dispositions du premier alinéa de l'article L. 121-1 ne sont pas applicables au conducteur, pour les infractions résultant d'une manœuvre d'un véhicule dont les fonctions de conduite sont déléguées à un système de conduite automatisé, lorsque ce système exerce, au moment des faits et dans les conditions prévues au I de l'article L. 319-3, le contrôle dynamique du véhicule.*
Le conducteur doit se tenir constamment en état et en position de répondre à une demande de reprise en main du système de conduite automatisé.
Les dispositions du premier alinéa de l'article L. 121-1 sont à nouveau applicables:
1° Dès l'instant où le conducteur exerce le contrôle dynamique du véhicule à la suite d'une reprise en main de celui-ci;
2° En l'absence de reprise en main du véhicule par le conducteur à l'issue de la période de transition faisant suite à une demande du système de conduite automatisé dans les conditions prévues au II de l'article L. 319-3;
3° Au conducteur qui ne respecte pas les sommations, injonctions ou indications données par les forces de l'ordre ou les règles de priorité de passage des véhicules d'intérêt général prioritaires prévues au présent code".

[1041] The definition can be found in the Decree of 29 June 2021 n. 2021-873, TRAT2034544 , at article 2: it entails "[t]he performance of all real-time operational and tactical functions required to move the vehicle", including "control of the lateral and longitudinal movement of the vehicle, monitoring of the road environment, response to events in road traffic, and preparation and reporting of maneuvers". See M. Giuca, "Disciplinare l'intelligenza artificiale. La riforma francese sulla responsabilità penale da uso di auto a guida autonoma", *Archivio Penale*, Vol. 2, 2022, p. 22.

[1042] "*Art. L. 319-3.-I. La décision d'activer un système de conduite automatisé est prise par le conducteur, préalablement informé par le système que ce dernier est en capacité d'exercer le contrôle dynamique du véhicule conformément à ses conditions d'utilisation.*

When the request of a takeover fails, or in the event of a serious failure, the ADS must also be able to put the vehicle in safety. Thus, according to French law, the AI system is seen as the focal point of responsibility, as it is tasked both with notifying drivers of its ability to exert dynamic control at specific intervals during the journey, as well as with alerting them of its inability to do so at other intervals, through the use of a "*demand de reprise*" notification.

The third paragraph of art. L.123-1 regulates the re-expansion of the scope of liability for the driver: it stipulates that the driver who either retakes dynamic control of the vehicle or fails to respond to a *demand de reprise* after a "transition period" (which remains undetermined) will be subject again to the provision of article L. 121.1.[1043]

Finally, in reference to the manufacturer of the vehicle, Article L.123-2 stipulates that the producer shall be held accountable for any unintentional injury or harm caused to an individual's life or physical well-being committed by the vehicle during intervals when the ADS was actively exerting control over the vehicle, as per its intended usage conditions, provided that a culpable act as defined by Article 121-3 of the French Penal Code can be established.

D) England and Scotland

At the end of January 2022, the Law Commission of England and Wales and the Scottish Law Commission ("Law Commissions") released a Joint Report ("Joint Report") on Automated Vehicles.

The aim of the Joint Report is to facilitate the implementation of specialized legislation, specifically, the Automated Vehicle Act. Its publication marks the first instance in which the

II.-Lorsque son état de fonctionnement ne lui permet plus d'exercer le contrôle dynamique du véhicule ou dès lors que les conditions d'utilisation ne sont plus remplies ou qu'il anticipe que ses conditions d'utilisation ne seront vraisemblablement plus remplies pendant l'exécution de la manœuvre, le système de conduite automatisé doit:
1° Alerter le conducteur;
2° Effectuer une demande de reprise en main;
3° Engager et exécuter une manœuvre à risque minimal à défaut de reprise en main à l'issue de la période de transition ou en cas de défaillance grave".

[1043] Giuca argues that this newly introduced immunity clause works as a mere *reconnaissance* of a conclusion which could have been reached by applying standard principles of criminal law, specifically the rules on negligence. Giuca, 2022, p. 29.

Law Commissions have been tasked with developing a legal framework *prior* to the emergence of future technological advancements.[1044].

The Joint Report defines an AV as a vehicle that is designed to be capable of driving itself ("self-driving vehicles").[1045] Self-driving vehicles operate in such a way that they do not need to be controlled and monitored by an individual, for at least a portion of a journey. Let us analyze the most relevant provisions of the proposal.

To begin with, the UK Law Commission propose the introduction of a new and independent authorization scheme[1046] for evaluating whether an ADS feature can be considered as self-driving according to the law or not.[1047] The attribution of the label "self-driving" to an ADS feature[1048] is relevant as it represents the prerequisite for the activation of the immunity clause. Indeed, once the ADS feature is correctly engaged, the human in the driving seat would acquire by law the new role of "user-in-charge", causing a change in the allocation of liability, as it will be discussed below.

By adopting the term "self-driving" the Law Commissions purposedly chose to take distance from the terms used in the SAE Taxonomy. [1049] The term is so cogent that the

---

[1044] Law Commissions Report, 2022, p. 1, para 1.1.

[1045] Ivi, p. 2, para 1.10.

[1046] Authorization would consist of a separate procedure from domestic, European or international approvals, which instead regard whether the vehicle can be placed on the market. Each ADS feature would have to be assessed. The objects of the authorization are three. First, the authorization authority must assess whether the feature reaches the legal threshold to be labelled as self-driving (Law Commissions Report, 2022, p. 27, para 2.57). Second, each ADS feature must be able to control the vehicle in a legal and safe way, even if the human user is not monitoring the driving environment, the vehicle, or the way the way the vehicle drives (Law Commissions Report, 2022, p. 67, para 4.66). The Law Commissions recommend that the new Automated Vehicle Act "require the Secretary of State for Transport to publish a safety standard against which the safety of automated driving can be measured" which "should include a comparison with harm caused by human drivers in Great Britain" (Recommendation 6, p. 27, para. 4.66). Measuring the performance of the AVs against those of human drivers would ensure public acceptance ("When deaths and injuries occur, it will be important to reassure the public that AVs are nevertheless safer than human drivers, and to have the evidence to support this claim", p. 66, para. 4.62). Third, the authorization authority will evaluate whether the ADS entity (ADSE) has sufficient resources to keep the vehicle updated and compliant with traffic laws in Great Britain and to deal with any kind of issue that might arise.

[1047] Law Commissions Report, 2022, Ch.2, pp. 69 ff.

[1048] An ADS "feature" is defined as "a combination of software and hardware which allows a vehicle to drive itself in a particular operational design domain (such as a motorway). Ivi, p. 135, note 239.

[1049] It is not a mere linguistic choice: the term was adopted in order to convey a legal boundary as opposed to a technical one. As will be discussed later, once this threshold is met, the individual occupying the

drafters recommend that it becomes "protected", in the sense of being safeguarded by two specific criminal offenses: first, "Describing unauthorised driving automation as 'self-driving"';[1050] second, "Misleading drivers that a vehicle does not need to be monitored".[1051]

Most importantly, the Law Commissions argue that they aim to "*draw a bright line*":[1052] the criminal liability of the person sitting in the driving seat of a self-driving vehicle shall be excluded for *any* harm arising from the *dynamic* driving task, in cases where the offense is committed by a vehicle which was previously authorized to deploy self-driving features, assuming that those features were properly engaged. In order to draw this "bright line", the recommendations create three new legal actors: the user-in-charge, the Authorised Self-Driving-Entity (ASDE) and the No-User-In-Charge (NUIC) operator.

Starting from the first, the user-in-charge is defined as the human being sitting in the driver's seat while a self-driving feature is engaged. His main role is "to take over driving, either following a transition demand or because of conscious choice".[1053] The user-in-charge enjoys immunity from "driving offences",[1054] provided that she has engaged the ADS correctly and that he has not tampered with the system.[1055] Driving offenses do not constitute a pre-existing category of crimes in UK legislation. They are defined in Joint Report as any offence involving "a breach of duty to monitor the driving environment and respond appropriately by using the vehicle controls to steer, accelerate, brake, turn on lights or

---

driver's seat (referred to as the "user-in-charge") would no longer be held liable for any harms resulting from the dynamic driving task.

[1050] Law Commissions Report, 2022, p. 126, para, 7.21.

[1051] Ivi, p. 129, para. 7.38.

[1052] Law Commissions Report, 2022, 5.46. A dynamic driving task is defined as "the real-time operational and tactical functions required to operate a vehicle in on-road traffic. It includes steering, accelerating and braking together with object and event detection and response", Law Commissions Report, xviii.

[1053] Ivi, p. 55, para. 4.1.

[1054] The user-in-charge remains liable for non-dynamic offenses and is responsible for: "(1) Duties to carry insurance; (2) Duties to maintain the vehicle in a roadworthy condition; (3) Any parking offence which continues after the ADS feature is disengaged; (4) Duties following accidents to provide information and report accidents; (5) Duties to ensure that child passengers wear seatbelts; (6) Duties relating to loading; and (7) Strategic route planning, including duties to pay tolls and charges." Ivi, Recommendation 45, p. 154, para 8.103.

[1055] "Recommendation 44. While a relevant ADS feature is engaged, the user-in-charge should not be liable for any criminal offence or civil penalty which arises from dynamic driving. The immunity should not apply if the user-in-charge has taken steps to override or alter the system so as to engage the ADS when it is not designed to function. The immunity should cease if the user-in-charge deliberately interferes with the functioning of the ADS." Law Commissions Report, 2022, p.149, para 8.79.

indicate".[1056] Examples of dynamic driving offences are dangerous driving, careless driving and exceeding the speed limit.

The definition of user-in-charge can be divided into four elements. She must be:

(1) *an individual* (a human or "natural person", rather than an organization);
(2) *who is in the vehicle* (hence not standing nearby or in a remote operations center);
(3) *in position to operate the driving controls* (for current vehicle design this entails that she in the driving seat);
(4) *while an ADS feature requiring a user-in-charge is engaged* (an ADS feature is engaged when it is switched on and remains so until the individual takes control of the vehicle, the transition period ends or it switches off at the end of a journey.[1057]

The user-in-charge is no average (reasonable) agent: users-in-charge must be "qualified and fit to drive", alike "average drivers" who are liable for conducts such as unlicensed driving or driving under the influence of alcohol or drugs; moreover, users-in-charge must be "receptive to a transition demand" and comply with other "driver responsibilities", which include insuring the vehicle and reporting accidents.[1058]

The Joint Report distinguishes between the duties of *monitoring* and of *receptivity*. The first entails checking the driving environment, the vehicle, or the way it drives. An ADS feature can be considered as self-driving only if it excludes this duty. Consequently, the user-in-charge shall not comply with the duty of monitoring. The second, instead, entails being receptive to a transition demand, i.e., the request by the vehicle for the human user to take over the dynamic driving. Think of the following example to understand the distinction: "[a] person who becomes aware of a fire alarm or a telephone ringing may not necessarily have been monitoring the fire alarm or the telephone".[1059]

[1056] Ivi, p. 146, para 8.62.
[1057] Law Commissions, "Automated Vehicles: Summary of joint report", Summary of LC Report No 404 / SLC Report No 258, HC 1068 SG/2022/15, 26 January 2022, p. 17, para 4.2. Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2022/01/AV-Summary-25-01-22-2.pdf.
[1058] Law Commissions Report, 2022, p. 21, para. 2.46.
[1059] SAE, 2021, p. 12.

Users-in-charge must comply with the duty of receptivity, i.e., they need to be receptive to a transition demand, which musts be communicated by clear, multi-sensory signals and give the user-in-charge sufficient time to gain situational awareness.[1060] We can find a duty of receptivity also in the French amendment, which provides that the driver shall constantly be in a condition and in a position to respond to a transition demand (*demande de reprise*) from the ADS.[1061] When it comes to transition demand and liability, time is of the essence: as soon as the transition period is over, the user-in-charge loses her immunity and she goes back to being legally treated as a driver. [1062] Yet, the Law Commissions do not specify how long this period should be, consequently leaving a major gap in the proposed regulation. Notably, in order for a user-in-charge to be receptive, she needs to know what she must be receptive to. Furthermore, she might need to "rehearse how to respond appropriately if the stimulus arises. That is why, in addition to installing fire alarms, organisations have fire drill".[1063] A question which remains unanswered, it follows, is the one of how a normal driver shall become a "fit" user-in-charge.

As it was highlighted, the immunity clause's application is lifted in case the user-in-charge fails to respond to a transition demand. In these cases, the ADS "should carry out a sufficient risk mitigation manoeuvre. Regulators will need to consider what is sufficient, but we would expect that (at a minimum) the vehicle should come to a controlled stop in lane with its hazard lights flashing".[1064] In terms of liability of the user-in-charge (now "normal" driver), the recommendations only state that "the law should impose consequences on a user-in-charge who fails to take over", without providing any kind of further instructions.[1065] Again, a gap occurs.

As a final point, it is relevant to focus briefly on the second and the third legal roles introduced by the Joint Report: the ADSE and the NUIC operators. These subjects are legal

---

[1060] Law Commissions Report, 2022, p. 13, para. 2.15.

[1061] "Le conducteur doit se tenir constamment en état et en position de répondre à une demande de reprise en main du système de conduite automatisé.". Art. L. 123-1, *Code de la route.*

[1062] "The length of the period is legally significant". Law Commissions Report, p. 40, para 3.27.

[1063] "Automated Vehicles: Consultation Paper 3 – A regulatory framework for automated vehicles A joint consultation paper", Law Commission Consultation Paper No 252, Scottish Law Commission Discussion Paper No 17, para 4.27, p. 41. Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2021/01/AV-CP3.pdf.

[1064] Law Commissions Report, 2022, p. 43, para 3.42.

[1065] Ivi, p. 160, para 8.132.

persons, rather than natural persons, and might coincide in cases where the vehicle manufacturer or developer is also the one providing a passenger service. An ASDE is defined as "the entity that puts an AV forward for authorisation as having self-driving features. It may be the vehicle manufacturer, or a software designer, or a joint venture between the two".[1066] The ASDE will have the duty to prove in the authorization process that the user-in-charge has sufficient time to gain situational awareness in cases of a transition demand and, if she fails to respond, that the vehicle is capable of sufficient mitigation against the risk of a crash.[1067] Other duties of the ASDE include duties relating to safety (such as ensuring that the vehicle continues to drive safely and in accordance with road rules)[1068] and duties of disclosure.[1069] The NUIC operator is a licensed legal person which oversees vehicles possessing a NUIC feature. While on a vehicle deploying NUIC features, a whole journey can be completed without any kind of intervention by a human on board. This does not mean that there would be no human being on board, but that when the ADS feature is of NUIC nature, any human in the car will be considered as a mere passenger. The NUIC operators need to have "oversight" of the vehicle any time a NUIC feature is engaged on a road or in another public place: they are "expected to respond to alerts from the vehicle if it encounters a problem it cannot deal with, or if it is involved in a collision", but they are not expected to monitor the driving environment.[1070] Oversight duties would include both remote assistance[1071] (for example, if the ADS detects an object in its lane which is too large to avoid and stops, remote assistance could imply providing instructions to the vehicle on how to deal with the obstruction) and fleet operations (for example, dealing with law-enforcement agencies or paying toll for the route).[1072] As stated in Recommendation 56, the regulator shall have powers to impose only *regulatory* sanction (such as warnings, civil penalties, suspension or withdrawal of license) to NUIC operators, hence they could not be

[1066] Law Commissions Report, 2022, p. 20, para 2.41.
[1067] Ivi, p. 81, para 5.65.
[1068] Law Commissions Report, 2022, pp. 85-86, para 5.95 (1).
[1069] Ivi, p. 86, para 5.96.
[1070] Law Commissions Report, 2022, p. 21, para 2.48.
[1071] Defined as "Event-driven provision, by a remotely located human, of information or advice to an ADS-equipped vehicle in driverless operation in order to facilitate trip continuation when the ADS encounters a situation it cannot manage". SAE Taxonomy J3016, 2021, para 3.23.
[1072] Law Commissions Report, p. 167, para. 9.14.

subjected to criminal liability.[1073] Additionally, the individual staff of the NUIC involved in remote driving of the vehicle "will face the same criminal liabilities as other drivers", for example if they are not trained or qualified enough.[1074] Yet, the Joint Report does not advise in favor on introducing new criminal offences relating to individual assistants.

E) Germany

In May 2021, the German Bundesrat passed a law which amended the German Road Traffic Act (*Straßenverkehrsgesetz* - StVG).[1075] Even though it does not directly regulate matters of criminal liability, it is of utmost interest for this discussion, as it attempts at providing a legal solution to moral dilemmas similar to those examined in the MIT's Moral machine Experiment. As such, Germany "paved the way to legislating higher levels of autonomous driving while international initiatives have been stalling".[1076]

As it will be highlighted, the connection between moral issues and issues of criminal liability is rendered evident also by the fact Germany's new law was based on the Report of the Ethics Commission on Automated and Connected Driving,[1077] which included amongst his members prof. Eric Hilgendorf, professor of criminal law and author/editor of numerous publication on the matter of automated driving and criminal law.[1078]

[1073] Ivi, p. 168, para. 9.120.

[1074] Law Commissions Report, p. 172, para. 9.42.

[1075] Bundestag, Gesetz zur Änderung des Straßenverkehrsgesetzes, 2021

[1076] A. Kriebitz, R. Max & C. Lütge, "The German Act on Autonomous Driving: Why Ethics Still Matter", *Philosophy & Technology*, Vol. 35, 2022.p. 11.

[1077] See also European Commission, Directorate-General for Research and Innovation, Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659), "Ethics of connected and automated vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility", Publications Office of the European Union, Luxembourg, 2020. Available at: https://data.europa.eu/doi/10.2777/035239.

[1078] See *ex multis*: E. Hilgendorf, S. Hötitzsch & L. Lutz (Eds.), *Rechtliche Aspekte automatisierter Fahrzeuge. Beiträge der 2. Würzburger Tagung zum Technikrecht im Oktober 2014*, Nomos, 2015; E. Hilgendorf (Ed.), *Autonome Systeme und neue Mobilität. Ausgewählte Beiträge zur 3. und 4. Würzburger Tagung zum Technikrech*. Nomos, 2017; E. Hilgendorf & J. Feldle (Eds.), *Digitization and the Law*, Nomos, 2018; E. Hilgendorf, "Autonome Systeme, künstliche Intelligenz und Roboter: Eine Orientierung aus strafrechtlicher Perspektive", in S. Barton (Ed.), *Festschrift für Thomas Fischer*, C.H. Beck, 2018; E. Hilgendorf, "Dilemma-Probleme beim automatisierten Fahren. Ein Beitrag zum Problem des Verrechnungsverbots im Zeitalter der Digitalisierung", *ZIS*, Vol. 130, No. 3, 2018.

The most interesting provision is the one at article 1 § 1e (2)2, which lists the technical requirements that must be complied with by a vehicle with autonomous driving functions (level 4 automation):

> (2) Vehicles with autonomous driving function must be equipped with a technical system that is capable of,
> [...]
> 2. autonomously comply with the traffic regulations addressed to the vehicle driver and be equipped with an accident-avoidance system which
>
> (a) Is designed to prevent and reduce damage,
>
> (b) in the event of *unavoidable alternative damage* to different legal goods, *takes into account the importance of the legal goods*, provided that the protection of human life has the highest priority, and
>
> (c) in the event of an *unavoidable alternative harm to human life*, does not factor in personal characteristics.[1079]

Admittedly, the law, specifically art. 1§ 1e (2)2 (c), "does not illuminate what personal characteristics are, nor does it state whether autonomous vehicles might hit individuals violating traffic rules in unavoidable accidents".[1080] These questions were addressed in the MIT's Moral Machine experiment, which was discussed above.[1081] A list of the aforementioned personal characteristics can be found in the principles elaborated by the Ethics Commission on Automated and Connected Driving, specifically at rule 9 (age, gender, physical or mental constitution). Actually, the law of May 2021 codifies part of the principles

---

[1079] Translated from German: *(2) Kraftfahrzeuge mit autonomer Fahrfunktion müssen über eine technische Ausrüstung verfügen, die in der Lage ist, […] 2. selbstständig den an die Fahrzeugführung gerichteten Verkehrsvorschriften zu entsprechen und die über ein System der Unfallvermeidung verfügt, das*

a) auf Schadensvermeidung und Schadensreduzierung ausgelegt ist,

b) bei einer unvermeidbaren alternativen Schädigung unterschiedlicher Rechtsgüter die Bedeutung der Rechtsgüter berücksichtigt, wobei der Schutz menschlichen Lebens die höchste Priorität besitzt, und

c) für den Fall einer unvermeidbaren alternativen Gefährdung von Menschenleben keine weitere Gewichtung anhand persönlicher Merkmale vorsieht".

Bundestag, Gesetz zur Änderung des Straßenverkehrsgesetzes, 202, Art. 1 § 1e.

[1080] Kriebitz, Max & Lütge, 2022, p. 9.

[1081] See Para. 5.3.2.

contained in the Report published in 2017 by the Ethics Commission, hence it will be analyzed in the following paragraphs.[1082]

In the Report, the Ethics Commission laid down twenty rules for automated and connected vehicular traffic. Amongst them, rule 2 states that the protection of individuals shall take precedence over all other "utilitarian considerations"[1083] and that the purpose of AVs should be "to reduce the level of harm until it is completely prevented".[1084] As a final point, it upholds that "[t]he licensing of automated systems is not justifiable unless it promises to produce at least a diminution in harm compared with human driving, in other words a *positive balance of risks"*.[1085] Moreover, according to rule 5, AVs should "*prevent accidents* wherever this is practically possible" and, based on the state of the art, "*the technology must be designed in such a way that critical situations"*,[1086] including moral dilemmas*, "do not arise in the first place*".[1087]

Moral dilemmas[1088] represent attractive hypothetical scenarios: they are able to stir up a discussion amongst any kind of audience. Their attractiveness lies in the fact that "they help to illustrate the basic values of a legal culture, in a manner which is also accessible to a broader public".[1089] The solutions proposed to solve moral dilemmas often "take on the

---

[1082] The Ethics Commission was appointed by the German Federal Minister of Transport and Digital Infrastructure (*Bundesministerium für Digitales und Verkehr*).

[1083] Ethics Commission Automated and Connected Driving, "Report", 2017, p. 10 ("Ethics Commission Report"). Available at: https://bmdv.bund.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile.

[1084] Ivi, p. 10.

[1085] Ethics Commission Report, 2017, p. 10.

[1086] Ibid.

[1087] Ethics Commission Report, 2017, p. 11.

[1088] Probably the most famous moral dilemma is the "trolley problem". Imagine that you are the driver of a trolley which is headed towards five track workmen who are repairing the track. In order to avoid running over the five track workmen, you must stop the trolley, but you realize that the brakes don't work. You see a side track on the right, on which there is only one track workmen. You are now presented with two choices: turn the trolley to the right, and kill one man, or keep the trolley going straight, and kill five. The invention of this type of moral dilemma is credited to philosopher Philippa Foot. The term "Trolley Problem" was coined later by Judith Jarvis Thomson. See P. Foot, "The Problem of Abortion and the Doctrine of the Double Effect", *Oxford Review,* 1978; J. J. Thomson "The Trolley Problem" *The Yale Law Journal,* Vol. 94, No. 66, 1985, pp. 1395-1415. Cf. Hilgendorf, "The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems", 2018, p. 60. The author mentions the decision of the German Federal Constitutional Court on the Aviation Security Act (*Luftsicherheitsgesetz*) of 2005 (BVerfGE, 15.2.2006 - 1 BvR 357/05).

[1089] Hilgendorf, "The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems", 2018, p. 57.

character of legal and social policy decisions"[1090] and turn into a "classic" "German culture (Kant)" vs British "utilitarian merchant spirit" (Bentham)[1091] debate.

As an example of the first approach, authors mention the decision of the German Federal Constitutional Court on the Aviation Security Act (*Luftsicherheitsgesetz*) of 2005.[1092] The decision regarded §14(3) of the Aviation Security Act, i.e., "whether a commercial airliner filled with innocent passengers, which had been hijacked by terrorists with the intent of using it as a weapon of mass destruction, for example, by crashing it into a city centre, could be shot down".[1093] The Court answered negatively, affirming that weighing lives would entail a violation of Article 1(1) of the *Grundgesetz*, as any individual life "in and of itself represents the highest possible maximum value"[1094] and therefore must be treated as a "non-balanceable".[1095]

With the development of AVs, and of the necessary discussion on their regulation, so came a conversation on how an AV should behave in a road traffic accident. Notably, one can notice a "*compulsion to analyze and to explicate*" [1096] when it comes to the developments of algorithms, which led to the analysis of incident scenarios from both an ethical and legal perspective. In Ch. 4, we discussed issues of alignment (i.e., the fact that AI systems might have different values from us) and of interpretation (i.e., the fact that moral rules are ambiguous) in connection to the development of moral machines. The inventors of the Moral Machine themselves refer to the workings of the German Ethics Commission and claim that their work "represents the first and only attempt so far to provide official guidelines for ethical choices of autonomous vehicles".[1097]

AV-based moral dilemmas, similarly to the PCM experiment,[1098] represent a further elaboration of these issues. For example, every German citizen might agree on the fact that

---

[1090] Ibid.

[1091] Ibid.

[1092] BVerfGE, 15.2.2006 - 1 BvR 357/05.

[1093] Hilgendorf, "The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems", 2018, p. 60.

[1094] Ivi, p. 65.

[1095] Hilgendorf, "The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems", 2018, p. 66.

[1096] Ivi, p. 59.

[1097] Awad & al., 2018, p. 60.

[1098] Discussed in Hilgendorf, "The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems", 2018, p. 64.

being inside or outside the Bavaria region is not an acceptable criterion for an AV to decide – in an accident situation – whether to choose the course of action that leads to killing the lowest number of people as opposed to the one that leads to killing the highest one (a so-called "Bavaria friendly collision algorithm").[1099] However, not every German citizen might agree on which criteria should be applied when the AV must decide between two options that lead to the death of the *same* amount of people: should it spare all the female innocent bystanders and kill all the male ones? In other words, this type of moral dilemmas prompts the question of how to break down into computable form the choices that an AI system must make in order to minimize the number of lives lost, if possible.

An attempt at regulating such choices can be found in rule 7 and rule 8, which address cases of "unavoidable hazardous situations", i.e., moral dilemmas. Rule 7 establishes that, in dilemma situations, the protection of human life must enjoy unconditional top priority when balanced with other legally protected interests.[1100] Accordingly, "within the constraints of what is technologically feasible, the systems must be programmed to *accept damage to animals or property* in a conflict if this means that personal injury can be prevented".[1101]

The solution seems straightforward, at first glance. Still, think of the case where the damage to property – which, in abstract, should not be preferred to the protection of life – amounts to causing oil spill from a road tanker or, even worse, the collapse of the power grid of an entire metropolitan area.[1102] Indeed, the first part of rule 8 recognizes that the "normalization" [1103] of moral dilemmas, i.e., the invention of generalized "lesser of two evil" solutions, is not feasible:

> Genuine dilemmatic decisions, such as a decision between one human life and another, depend on the actual specific situation, incorporating "unpredictable" behaviour by parties affected. They can thus *not be clearly standardized*, *nor can they be programmed such that they are ethically unquestionable*. Technological systems must be designed to avoid accidents. However, *they cannot be standardized to a complex or intuitive assessment* of the impacts of an accident in such a way that they can *replace or anticipate the decision of a*

[1099] Ibid.
[1100] Ethics Commission Report, 2017, p. 11.
[1101] Ibid.
[1102] The example is made by the Ethics Commission at p. 17.
[1103] Ethics Commission Report, 2017, p. 17.

> *responsible driver with the moral capacity to make correct judgements*.[1104] It is true that a human driver would be acting unlawfully if he killed a person in an emergency to save the lives of one or more other persons, but he would not necessarily be acting *culpably*. Such *legal* judgements, made in retrospect and taking special circumstances into account, *cannot readily be transformed into abstract/general ex ante appraisals* and thus also not into corresponding programming activities. For this reason, perhaps more than any other, it would be desirable for an independent public sector agency (for instance a Federal Bureau for the Investigation of Accidents Involving Automated Transport Systems or a Federal Office for Safety in Automated and Connected Transport) to systematically process the lessons learned.[1105]

Nonetheless, Rule 9 does lay down instructions on which are summarized as "No selection of humans, no offsetting of victims, but principle of damage minimization":

> In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties.

By allowing for an AI system which can be programmed as to reduce the number of personal injuries, Rule 9 departs from the principle enshrined in the Federal Constitutional Court's judgment on the Aviation Security Act. Indeed, the application of said principle would most likely lead to programming the AV not to do anything: perhaps it is no coincidence that in the MIT's Moral Machine German nationals scored 7th worldwide on "preferring inaction".[1106]

According to the drafters of the Ethical Guidelines, in the case of AVs the difference stands in the fact that

---

[1104] Ibid.

[1105] Ethics Commission Report, 2017, p. 11.

[1106] Moreover, Rule 9 clashes with the findings of the MIT's Moral Machine, which discovered strong preference for sparing the lives of the young. Awad et al., 2018, p. 60.

> a probability forecast has to be made from out of the situation, in which the identity of the injured or killed parties is not yet known […] Programming to minimize the number of victims (damage to property to take precedence over personal injury, personal injury to take precedence over death, lowest possible number of persons injured or killed) could thus be justified, at any rate without breaching Article 1(1) of the Basic Law, if the programming reduced the risk to every single road user in equal measure.[1107]

The approach, it is argued, is similar to the one adopted with immunizations where "statutorily imposed compulsory vaccination results in a general minimization of the risk without it being known beforehand whether the vaccinated person will belong to the group of the (few) harmed (sacrificed) parties" and, despite this, "it is in the interests of everyone to be vaccinated and reduce the overall risk of infection".[1108]

Conclusively, though, the Ethics Commission claims that it "refuses to infer from this that the lives of humans can be 'offset' against those of other humans in emergency situations so that it could be permissible to sacrifice one person in order to save several others".[1109] Therefore, the decision to save more as many lives as possible would be allowed only "if several lives are already immediately threatened".[1110] The issue, at first glance, does not appear as specifically related to AI systems. Rather, it seems to belong to the general category of ethical problems. Conclusively, the Ethics Commission states that, as recognized by the Commission itself, it did not reach a "satisfactory end"[1111] on the matter.

Let us conclude with a reflection connected to liability. Imagine that two seriously injured people (A and B) are lying on the road and an AV, which is carrying a human passenger C, is approaching at high speed. If the AV skews right to avoid running over (and killing) A and B, it will hit D and E, who are standing at the side of the road. Based on (very fast-paced) calculations, including an unspecified risk of injuring C, the almost certain probability of injuring D and E, and the likely risk of killing A and B, it decides not to swerve.

[1107] Ethics Commission Report, 2017, p. 18.
[1108] Ibid.
[1109] Ethics Commission Report, 2017, p. 18.
[1110] Ibid.
[1111] Ethics Commission Report, 2017, p. 18.

As a consequence, A and B die from the impact. Could it be argued that the AV's behavior is justified or excused?[1112]

Defenses work as "fixes" in criminal legal systems: they recalibrate the aim of punishment in a way as to not impose an unbearable burden on people.[1113] Seemingly, the situation descripted in the example could be reconducted to the realm of application of a necessity defense. Necessity implies that legal systems cannot predict and regulate beforehand all situations in which a balancing of protected legal interests is needed. Consequently, through necessity defenses we accept that there are situations "in which the offence definition is fulfilled, but the defendant should not be held liable, because all things considered, he did the right thing for society".[1114] Let us assume now that the AV is capable of calculating the probability of injury of A, B, C, D, and E, and that human lives could be subjected to a balancing act. Consequently, it could be argued that the choice of the AV not to swerve amounts to choosing the *lesser evil* outcome.[1115] Thus, is uncertain at this point how the abovementioned "probability forecast" would be pursued. The AV should be able to calculate exactly, in the concrete and urgent situation (in a matter of seconds), which is the probability of injury for every person present in the scene of the accident. According to Hilgendorf, in the foreseeable future "[a]t best it will be possible to make qualitative or comparative statements, i.e. statements such as 'almost certain', 'very likely', 'very unlikely', or statements such as 'event A is more likely than event B'".[1116]

One could also contemplate whether introducing an excuse similar to duress would be a preferrable option.[1117] In situations like the moral dilemmas outlined above, the best solution might be to accept that the AV system simply *could not have done otherwise*, as the choice between maintaining a straight course and swerving right, i.e., between killing A and B or D and E, was not a real choice. Indeed, if it is a widely accepted principle that "[t]he law cannot

---

[1112] The example is adapted from Hilgendorf, "The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems", 2018, p. 75.

[1113] Blomsma, 2012, p. 321.

[1114] Ivi, p. 372.

[1115] Lagioia & Sartor, 2020, p. 6.

[1116] Hilgendorf, "The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems", 2018, p. 75.

[1117] Apparently, the difference between necessity and duress is only "gradual" since they "both revolve around the infringement of the freedom of choice of the actor". They can be distinguished based on "the nature of the pressure to commit the offence" which in necessity "arises from the necessity to choose", where instead in duress it comes from the "danger (to life or limb)". Blomsma, 2012, p. 370.

ask of anyone more than they can be legitimately expected to do, either intellectually or emotionally", why would we expect something different of AI systems?

The argument becomes even more compelling if one takes into consideration the problems of pinning blameworthiness on AI systems. The application of an excuse would also bring forth positive effects from a communicative perspective, since it would not amount to saying that the killings were legitimate (as it would be in the case of a justification), but rather to claiming that the conduct was not blameworthy.[1118] Indeed, "[t]he defendant who raises an excuse confirms the wrongdoing and merely seeks to assert an explanation for his conduct" [1119] and "the acceptance of an excuse makes clear to the defendant and the public that what he did was wrong".[1120] Conclusively, the application of defenses to AI systems, as it was already argued with regards to the insanity of defense,[1121] is a path of investigation which definitely deserves attention for future research.

7.4 CONCLUSIONS

Chapter 3 underlined how most of the academic discourse is polarized by discussing the relationship between law and ethics. The trend is also visible when analyzing regulation on the mater, as most proposals for AI regulation aim at ensuring "accountability" and affirming soft law principles directed at the development of ethical or trustworthy AI, rather than the development of hard law instruments. Accountability is one of seven principles mentioned by the AI-HLEG[1122] established by the European Commission in its "Ethical Guidelines" and is a recurring element in European regulatory initiatives.[1123] According to some, this use of ethics is problematic, as it would consist to nothing more than a display of a "law

[1118] Ivi, p. 325.
[1119] Blomsma, 2012, p. 325.
[1120] Ibid.
[1121] See Ch. 5 at Para. 4.1.
[1122] AI-HLEG, 2019.
[1123] Resolution of the European Parliament, *Framework on the Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies*, 2020/2012(INL), Oct. 20, 2020, arts 9, 22, 23, 72, 96, 102.

conception of ethics",[1124] i.e., "a view on the ethics endeavour that makes it a sort of replica of law".[1125]

The case studies which were analyzed in this Chapter represent, in some way, an exception to this trend, as they make a step further and try to develop principles which could be applied to the ascription of criminal liability connected to AI systems. Germany, discussed at E, develops arguments related to the intersection of ethical principles and criminal law principles in the specific field of AV systems. Regrettably, none of the case studies focuses extensively on the liability *of* AI systems and rarely refers to it. Nevertheless, they are valuable inputs to this research, as they are proofs that matters of AI and criminal liability are not relevant just for Hollywood movies and criminal law scholars.

[1124] G.E.M. Anscombe, "Modern moral philosophy" *Philosophy,* Vol. 33, No. 124, 1968, pp. 1-19.
[1125] A. Rességuier & R. Rodrigues, "'AI ethics should not remain toothless! A call to bring back the teeth of ethics", *Big Data & Society*, 2020.

8 CONCLUSIONS

But epers are no more transient legal entities.
Recall that corporations and organizations of all sorts come and go; they rise and fall like Italian governments.

Curtis E.A. Karnow, 'The Encrypted Self: Fleshing Out the Rights of Electronic Personalities', Marshall J. Computer & Info. L. 1 (1994)

8.1 Incipit – **8.2** The Definitional Question – **8.3** A *mare magnum* of Scholarly Literature. The Introspective Question – **8.4** Anthropocentrism and Responsibility of Machines. The Attribution Question – **8.4.1** Holding AI to a (Higher) Moral Standard – **8.4.2** Retributivists at Heart? – **8.4.3** Re-Evaluating the AI-corporation comparison – **8.4.4** Epicentres of Liability: the Human Culprit – **8.5** General Conclusions. The Backward and Future-Facing Question – **8.5.1** Looking Forward – **8.5.2** Looking Backwards.

8.1 INCIPIT

One can identify a very clear parabola for this research. It started with "machines going rogue",[1126] factories transforming humans into paperclips, and other apocalyptic scenarios – only to find out that the biggest threats might be toasters and airplanes.[1127] What happened? Were our first assumptions wrong? Not at all: criminal law, starting with the industrialization period, has had to adapt its imputation models to technological progress.[1128] The "AI-revolution", be it toasters or evil robots, is another episode of this "saga".

Apropos, one of the purposes of this research was to identify and evaluate which are the novelties – if any – that were brought by the incurrence of AI and what was their impact upon the classical constructs of criminal law. Coming to an end of the study, it can be

[1126] Badea & Artus, 2022, pp. 1-2.
[1127] "One example of AI causing harm through an attempt to carry out its stipulated goal is the intelligent toaster which burns a house down in a quest to make as much toast as possible". Turner, 2019, p. 204.
[1128] M. Simmler, "Automation", in P. Caeiro et al. (Eds.), *Elgar Encyclopedia of Crime and Criminal Justice*, Vol. 1, 2023.

concluded that AI systems indisputably have the *capacity to act* in a way that can be generally considered by society as *offensive* to protected legal goods, i.e., that they can act like "criminals". This calls for the necessity of society to answer, somehow.

The present conclusive Chapter will retrace the questions, and the interim conclusions, which were posed throughout the thesis and identify avenues for further inquiry. The main research question was: *to what extent is a theoretical framework of criminal liability for non-human agents needed and feasible?* Such a question prompted a set of sub-questions, which were addressed in the different Chapters.[1129] This conclusive Chapter will mirror the general structure of the thesis and build upon the interim findings disseminated throughout the research.

Thus, besides rearticulating the arguments which were already put forth, I will add fresh perspectives and propose new directions for future inquiries. To do so, the avenues of this research were redefined according to four main issues, which will be addressed as follows. Paragraph 8.2 will expand upon the question of defining AI and its repercussions on criminal law (the *definitional* question). In Paragraph 8.3 I will situate my research in the realm of the scholarly literature on AI and criminal law analyzed in Chapter 3 (the *introspective* question). Paragraph 8.4 will discuss the attribution of criminal liability to AI systems through the lenses of anthropocentrism (the *attribution* question), specifically expanding on why we are prone to hold AI to higher moral standards (para. 8.4.1). It will then touch upon theories of retributivism and deterrence (para. 8.4.2) and the analogy between AI liability and corporate criminal liability (para. 8.4.3). Paragraph 8.4.4 will discuss the liability of humans *for* machines also in light of the findings of Chapter 7. Finally, building on the previous sections I will draw general conclusions for this study at Paragraph 8.5. (the *backward and future-facing* question). Specifically, I will return to the main research question (para 8.5), expand on questions for the future (para. 8.5.2), and outline the challenges which I encountered throughout this research (para 8.5.3).

[1129] These are: Chapter 2 – why is the issue of defining AI an issue? What are the basic functions of AI?; Chapter 3 – What is the state of the art of the scholarly debate on AI and criminal law? Which are the most recurrent questions and what are the answers? Which aspects are being neglected?; Chapter 4 – What does ascribing criminal liability to an AI system entail? Chapter 5 – Can an AI system be considered a criminal agent?; Chapter 6 – Can an AI system fulfil the *mens rea* and *actus reus* requirements of a criminal offense?; Chapter 7 – what is the state of the art on criminal law regulation on AI?.

8.2 THE DEFINITIONAL QUESTION

This research started with examples of AIs "going bad". Such a *in medias res* immediately gave an idea of the relevance of the topic, presented in a language which could be understood by criminal lawyers. In fact, I believe that AI and criminal law share a problem of miscommunication: criminal law does not understand what AI is, and vice versa. This research represents a significant step in bridging such a linguistic divide.

On a preliminary note, it must be mentioned that I gave little or no attention to AGI in this study, and I did so on purpose. Even if we were to assume that AGI will ever be created, discussing its liability would mandate for a different methodological approach, hence it was excluded from the scope of analysis to maintain the research's coherence. Indeed, the most striking aspect of (narrow) AI is probably that it is so smart it's stupid, yet it might commit a crime.[1130] It follows that whether a criminal system decided to attribute criminal liability to AI systems, it would have to do so based on a different paradigm of intelligence than the one of human beings, since it is not comparable to the intelligence displayed by AI systems.

There are two sides to the "definitional question". The *first* strict meaning of the definitional question is addressed in Chapter 2, which dealt with why the issue of defining AI is an issue and then explored the basic functioning of AI. The lack of a strict *technical* definition of AI clashes blatantly with the principle of legality, which rejects general and ambiguous concepts. Indeed, legality struggles with AI since it is a label which refers to an everchanging concept. Moreover, not only AI systems are not a definite entity, but they cannot establish themselves as one. Furthermore, there is no consensus on *who* should define AI (Computer scientists? Philosophers? Legal scholars?). Surely, as Chapter 3 shows, the lack of a universal definition of AI has not stopped legal scholars from approaching the topic of "AI law". According to some, "[a]lthough there is nothing clear so far, jurisprudence can conjure up its own definition of AI at least in relation to legal discussion".[1131] Apropos,

[1130] C. Quirk, "So Smart It's Stupid", *Alumni*, 1 April 2022. Available at: https://alumni.msu.edu/stay-informed/alumni-stories/so-smart-its-stupid-the-weirdness-of-ai.
[1131] Lee, 2021, pp. 311-312.

Chapter 2 served indeed this very purpose: it "conjured up"[1132] a definition of AI which could be used in the legal discussion carried out in this research, and not only there.

The *second* aspect of the definitional question is broader and relates to the concept of AI as an *agent* in the eyes of (criminal) law. While this aspect is addressed directly in Chapter 5, it also lingers throughout the whole research, working as a *fil conducteur* in the present work. As a matter of fact, not only there is no agreed-upon scientific definition of AI, but there is also no *legal* definition of which characteristics an AI system should possess in order to be qualified as capable of criminal action. In other words, the two aspects of the definitional question are entangled.

Legal personhood (i.e., being able to "act in law")[1133] is a purely artificial construct: it is, by definition, "attributed by positive law (statute or common law) and cannot be assumed".[1134] Furthermore, it is well-known that "legal personhood is already dissociated from the human substrate",[1135] such as in the case of corporations, hence, according to some, "there would be no way to deny that AI can be a legal person".[1136] Nonetheless, "just because it is possible, that does not mean it should be a good idea".[1137]

Legal agency means possessing the capabilities, which are attributed by law, "to act in law and to be liable for one's own actions".[1138] It can be attributed to subjects in different legal domains. The same subject can also be deemed an actor in a certain field (e.g., a corporation in private law) and not in others (e.g., a corporation in legal systems which do not accept corporate criminal liability).[1139] Certainly, "acting in criminal law" necessitates peculiar rules, since "criminal law liability with its emphasis on censure assumes a kind of moral agency that is not obvious in the case of current day autonomous systems".[1140]

Having acknowledged this, in Chapter 5.1. I hypothesized that this study would not answer the definitional question once and for all. Not only there is a lack of a common

---

[1132] Lee, 2021, pp. 311-312.
[1133] Hildebrandt, 2019, p. 241.
[1134] Ivi, p. 242.
[1135] S. M. C. Avila Negri, "Robots as Legal Persons: Electronic Personhood in Robotics and Artificial Intelligence", *Frontiers in Robotics and AI*, Vol. 8, 2021, p. 2.
[1136] Ibid.
[1137] Ibid.
[1138] Avila Negri, 2021, p. 243.
[1139] Hildebrandt, 2019, p. 240.
[1140] Ibid.

understanding on what AI is, in terms of which technical/physical attributes it should display (according the first connotation of the *definitional question*), but there is also a *lacuna* in discussing whether AI could be considered an actor in terms of criminal capacity. I argue here that the adoption of the AI-HLEG's definition of AI systems[1141] as the definition of reference turned out to be appropriate in addressing the definitional question for the purposes of this research.[1142]

As a matter of fact, such a definition, which is not tied to the concept of human intelligence, proved to possess enough "freedom of scope" to capture "the whole range of *[AI]* applications finding their way into practice today and in the near future".[1143] Moreover, since the definition of AI will necessarily change over time, I believe that the AI-HLEG's definition could work as a valuable starting point to elaborate a definition of AI systems suitable for the criminal law domain.

Surely, adopting a working definition made sure that the field of investigation of this research remained clear, as compared to the "blur" which sometimes characterized the works of the scholars discussed in Chapter 3 (who rarely addressed what they meant by "AI"). Ultimately, one of the added values of this research is that it implemented, and discussed, a definition of AI systems with the cognizance of the significance that definitions possess in criminal law.

Finally, this research represents a valuable input to the discussion on whether AI systems could be considered as *Rechtspersonen* for the purposes of criminal law. Differently from what happens when seeking a definition of AI, one can identify a solid consensus on the definition of criminal capacities (with a number of variations). By relying on such a

---

[1141] According to which AI systems are "systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal". AI-HLEG, 2018, p. 9.

[1142] I do not argue here that I found the ultimate solution to defining AI, rather that the solution adopted proved successfully to overcoming the challenges posed to this very research by the definitional question. I also argue that it could prove useful for further researchers and policy/law makers.

[1143] Sheikh, Prins & Schrijvers, 2023, p. 19. The authors further state that "Instead of considering AI as a discipline that can be clearly delineated, with uncomplicated definitions and fixed methodologies, it is more useful to see it as a complex and diverse field focused on a certain horizon. The dot on that horizon is the understanding and simulation of all human intellectual skills. This goal is also called 'artificial general intelligence' or AGI (other names are 'strong AI' and 'full AI'). However, it remains to be seen whether this dot, with such a generic definition of AI, will ever be reached". Ibid.

doctrinal background, as it will be further expanded upon in the following sections, I maintain that AI systems should be treated as agents in criminal law only if they were to display the capacity to be truly susceptible to the command of a criminal norm.

8.3 A *MARE MAGNUM* OF SCHOLARLY LITERATURE.[1144] THE INTROSPECTIVE QUESTION

In 2020, a group of researchers from Maastricht University conducted a study on the scope of legal literature on AI with the aid of a ML technique called topic modelling.[1145-1146] The researchers found that scholarly output boomed in the so-called "deep learning era".[1147] As the authors argue, "with over 2500 publications already by the year 2015 referring to 'artificial intelligence' […] it may no longer be realistic to assume that researchers can keep up with legal research on AI, or the number of publications in general".[1148]

The literature review conducted in Chapter 3 (which results from analyzing over 100 documents of scholarly literature on the topic of AI and criminal law written in three languages) is an effort to systematize part of this *mare magnum*[1149] of scholarly literature. Besides it being a recognition of the status quo in scholarly literature, it also prompted an introspective question, which will be answered at the end of this paragraph: where does this researcher situate herself in the debate?

Expecting new legal scholarship to be able to account for each and any publication available on AI-law would mean imposing a cumbersome burden on researchers, especially on those who are at the early stages of their career. It is argued here that legal scholars do

---

[1144] I developed part of these reflections in A. Giannini, "Artificial intelligence, ethics, law: a view on the Italian and American debate (and on their differences*)", Netherlands Journal of Legal Philosophy*, Vol. 51, No. 2, 2022, pp. 248-263.

[1145] Rosca et al., 2020.

[1146] Topic Modelling is a method for classifying collections of documents. The authors adopted Latent Dirichlet Allocation (LDA) Topic Modelling. They used the tool to identify recurring topics in 3931 journal articles on AI legal research.

[1147] The expression refers to the period which includes the early 2000s until today. Goanta et al., 2020, p. 331.

[1148] Rosca et al., 2020, p. 1.

[1149] "First coined to describe the ocean ring that enclosed the earth in ancient times, the term *mare magnum* today has a strictly negative metaphorical meaning referring to a vast mass that converts into chaos", P. Basso, "*Mare magnum"*, *Parolechiave,* Vol. 1, 2022, p.1.

not need to entertain a "gold rush" to who reads (or publishes) fastest on the topic: AI technology might be developing rapidly, but the legal tools to address this rapid development have been the same for the past 50-100 years or so. This is reflected by the fact that legal scholarship analyzed in Chapter 3 presents recurring traits. These traits are visible even amongst researchers whom have very different cultural and legal backgrounds.

In Chapter 3, scholars were classified according to three categories: expansionists, moderates, and skeptics. The categorization was based on how each author approached the general issue of criminal law and AI, specifically with regards to matters pertaining to the general part of criminal law. It follows that it does not include authors (or the reflections of the authors part of the literature review) who focused on the impact of AI from a criminal procedure perspective, nor reflections on the use of AI as a tool to commit crimes and its impact on cybercrimes.

Throughout the analysis several recurring questions were identified and compared, as summarized at Chapter 3.5.[1150] This led to the identification of some weaknesses/gaps in the debate.[1151] First and foremost, it must be highlighted that no one had ever conducted an analysis as the one contained in Chapter 3 before. As such, this research already represents a novelty.

Furthermore, it is now possible to come back to the individual gaps identified and assess whether this research addressed them. The answer is positive: first, the study did not discard Hallevy's theories readily, rather, it analyzed them in depth also in comparison to

---

[1150] The questions/topics identified were: I Are robots/AI systems capable of moral reasoning?; II What is the relevance of moral reasoning when ascribing criminal liability?; III Can an AI system "act"?; IV Can an AI system be culpable?; V Why should we punish AI systems and how should we punish AI systems?; VI Permitted risk and dangerous activities; VII Are there more feasible alternatives to applying criminal law (such as torts law or administrative sanctions)?; VIII What is the relationship between corporate criminal liability and AI criminal liability?; IX Should human agents be liable for the AI's misbehavior? ; X The creation of new negligence crimes and the application of existing negligence offenses so to humans-in-the-loop (how should foreseeability and reasonable care be defined?).

[1151] I Civil criminal legal scholars discard Hallevy's theories quite effortlessly. II No or little attempts at providing a definition of AI and of AI agent applicable to criminal law, no discussion of the impact of the lack of a definition on the matter; III No or little attempts at addressing the preliminary issue of considering AI as a subject of criminal law (i.e., criminal capacity of AI); IV No or little discussion of whether an AI system can act, think or want by criminal legal scholar; V No or little discussion of whether conduct of AI systems could be excused or justified by a defense; VI Confusion on identifying precisely who would be the responsible human agent; VII No or little discussion on how AI behavior interferes with theories of causality.

other legal scholars (gap I, addressed at Ch. 3); second, the study provided a definition of AI and of AI agent applicable to criminal law, while thoroughly discussing the impact of the lack of a definition on the matter (gap II, addressed at Ch.2, Ch. 5 and Ch. 6); third, the study tackled the issue of criminal capacity of AI systems (gap III, addressed at Ch. 5); fourth, as discussing the differences between the way human and AI systems can act, think, or want is an extremely vast topic, the matter was addressed merely from a criminal legal perspective, rather than from a philosophical one (gap IV, addressed at Ch. 5.3); fifth, the study discussed whether AI systems could be justified by an insanity defense (gap V, addressed at Ch.5.4); sixth, the research addressed the issue of understanding who the responsible human agents should be and the problem of data errors (gap VI, addressed at Ch. 6.2.3); finally, the research tackled how AI behavior interferes with theories of causality, specifically which factors could lead to "failures of causation" (gap VII, addressed at Ch. 6.3.2).

Classifying authors was, at times, challenging: first, besides certain specific cases, the categorization was not always clear cut: think for example of Gless-Silverman-Weigend (skeptics), who leave the option of criminal culpability of robots open in case in the future they were to become able to truly determine themselves on moral grounds[1152] or of Simmler-Markwalder (moderates), who strongly believe that robots will be addressees of normative expectations in the future.[1153] As the literature review represents the first attempt to be conducted on such a large scale, it leaves space for debate and improvement. An interesting development of Chapter 3 would be if it sparked a debate amongst the authors therein regarding whether they agree with the categorization or not.

It is worth mentioning, at this point, that I identify myself with the moderate school. The reasons vary: most importantly, I do not discard discussing criminal liability *of* AI systems outright. The sole fact of writing a PhD on the topic which focuses primarily on the liability of non-humans,[1154] and treats the liability of humans-behind-the-machine as an ancillary topic, determines the automatic exclusion from the category of skeptics, for now (it cannot be assumed that more extended "skeptic" writings will emerge in the future). This research

---

1152 Gless, Silverman & Weigend, 2016, p.423.

1153 Simmler & Markwalder, 2019.

1154 As Lagioia and Sartor argue, "Given the current socio-technical arrangements, we are not arguing for this approach [*punishing AI systems*], but it may be useful to start speculating about it". Lagioia & Sartor, 2020, pp. 27-28.

cannot be placed in the expansionist front either, even though it sympathizes with many of the ideas of its representatives.

Truth be told, one of the aims of this research was to at least plant a seed of doubt in the mind of the most ardent skeptics, leading them to think that one could actually build a decent a case for imputing criminal liability on AI systems. It surely worked with me, up to a certain point. Notwithstanding this, there is one moderate argument which – from the beginning – was more compelling than any other: when it comes to directly punishing AI systems it is not as much matter of feasibility, as it is of *unnecessity*. As of today, criminal liability of AI systems is not needed. Probably, such a statement makes me a weak moderate. It is not excluded that I will join the dark-side of the skeptics, sooner or later.

Ideally, Chapter 3 is meant as a tool to be consulted by all those interested to the intersection between AI and the general part of criminal law. As such, the expectation is that it will be of use as a general framework of reference for future AI legal scholars who are looking for a perspective on current (and future) paths of investigation. Moreover, Chapter 3 can work as a case study of the differences between common law and civil law scholars in approaching conflict, which in the present case is represented by the impact of AI technologies on our society. As a matter of fact, one could study the debate reconstructed at Ch. 3 as a form of expression of the civil/common law divide. Indeed, where the "common law mind"[1155] tends to find a convincing pragmatic solution, the civil law mind tries to solve the conflict beforehand "through hierarchic organized norms".[1156] Consequently, common law systems are traditionally more receptive to development than civil legal systems. This trend is visible amongst some of the authors of Ch. 3. Perhaps most of the "civilians", whose work was examined in the literature review, resort to such solutions since they come from extremely doctrinal legal cultures, such as the German one,[1157] or from criminal legal systems which are deeply rooted in a principle of strict legality and of "personal" criminal liability, such as Italy. Tentatively, the research tried to learn from both minds.

[1155] A. W. B. Simpson, *Legal Theory and Legal History: Essays on the Common Law*, The Hambledon Press, 1987, p.394.

[1156] Beck, 2010, p. 1105.

[1157] See R. Hofmann, "Formalism versus pragmatism – A comparative legal and empirical analysis of the German and Dutch criminal justice systems with regard to effectiveness and efficiency", *Maastricht Journal of European and Comparative Law Issue,* Vol. 28, No. 2, 2021, pp. 452-478; Keiler & Roef (Eds.), 2019.

Finally, what could be the role of a new legal scholar in this *mare magnum*? Is the field saturated? A positive outlook is desired. The AI legal scholar of the future could focus on specific problems (e.g., negligence, causality, defenses) and/or on specific sectors (e.g., driving automation, healthcare). This type of research would exploit existing literature on AI law, including the present research, to its fullest. Assuredly, since Chapter 3 reports Italian and German voices of the debate into English, it could be improved by new contributions of scholars outside the (linguistical) scope of analysis.

8.4 ANTHROPOCENTRISM AND RESPONSIBILITY OF MACHINES. THE ATTRIBUTION QUESTION

8.4.1 Holding AI to a Higher (Moral) Standard

Scientific research is – nowadays – based on Newtonian hermeneutics, where "reality is considered to be built from impersonal, inanimate matter, and causal relationships between them".[1158] It follows that "we no longer consider an earthquake today as an expression of 'anger' of some God, but as a causal chain of tectonic events leading to the catastrophe".[1159] The development of AI systems is also based on Newtonian hermeneutics. This makes applying typical "anthropomorphic constructs" to AI, such as "free-will, conscience, trust, desire, anger [*and suffering*]",[1160] cumbersome. Hence,

> "shaming" or expressing disapproval, disgust and anger against a reprehensible act can act as a *deterrent* to a human; but machines hitherto don't respond to such expressive rhetoric. While we can appeal to the *conscience* of a human wrong-doer to make them correct their actions, no such mechanism exist for interacting with AI that is about to do something irresponsible.[1161]

[1158] S. Srinivasa & J. Deshmukh, *AI and the Sense of Self*, Srinivasa, Srinath and Jayati Deshmukh. "AI and the Sense of Self." *ArXiv abs/2201.05576*, 2021, p. 2.
[1159] Ibid.
[1160] Ibid [emphasis added].
[1161] Ibid [emphasis added].

One should note, though, that it is debated – much as it was debated with regards to corporations – "whether human-like prerequisites for moral agency should be imposed on machines or if a hard line should be drawn between human moral agency and that of machines".[1162] Certainly, this is nothing new, since also with CCL scholars and lawmakers had to face the fact that concepts such as *mens rea* and *actus reus* "which make perfectly good sense when applied to individuals, do not translate easily to an *inanimate fictional entity*".[1163] Thus, it must be stressed that AI systems *are not inanimate fictional entities*: some of them "live" and operate in the digital and/or physical world. In this, AI systems differ profoundly from corporations.

Let us now focus, then, on the matter of translating human constructs to machines. As it was revealed in Chapter 5.3.2, researchers have been exploring the possibility of creating "moral machines", or "AMAs",[1164] for some time now. Specifically, they are trying to create machines that are capable of implementing ethical principles and moral decision-making, so that they behave towards human users and other machines in an ethically acceptable way.[1165]

Some of the difficulties of translating morality into code were already highlighted in Paragraph 5.3.2. Here, we only need to be reminded of the fact that designing AMAs, in a nutshell, entails "defining the moral behaviors or ethics that the system will follow, implementing such moral behaviors or ethics, and operationalizing them".[1166] In "Robot Criminals", discussed in Paragraph 3.2.2, Hu contends that a new criminal code for robots could establish a "minimum set of *moral standards* to which all smart robots must adhere".[1167] On a similar note, Germany established via law that AVs must be capable of prioritizing legal goods when faced with moral dilemmas.[1168] The question which arises then is: which moral rules should be taught to AMAs? In other words, "who would be the one to identify 'true

---

[1162] A. Martinho, "Perspectives about artificial moral agents", *AI and Ethics*, Vol. 1, 2021, p.482.

[1163] J. Gobert & M. Punch*, Rethinking Corporate Crime*, Cambridge University Press, 2003, p.10 [emphasis added].

[1164] Artificial Moral Agents.

[1165] Martinho, 2021, p. 481.

[1166] Ibid. When addressing these topics, one is prone to ask herself which moral rules should be taught to AI systems. Moreover, since "[e]thics has to be operationalized so that an AMA is able to recognize a moral situation, weigh up possible moral actions, make moral judgments, and execute them", it would be interesting to study also "how a machine moralizes". Ivi, p. 482.

[1167] Hu, 2019, p. 502,

[1168] See Ch.8, E.

morality'?".[1169] This is a question intrinsically connected to criminalization theories, specifically directed to those who belong to the "legal moralism" school of thought.[1170] It is questionable how these moral standards would be developed, since the human kind has not agreed yet on a universal set of moral rules for themselves, let alone for artificial beings.

Furthermore, one of the most prominent issues regarding the development of AMAs is the one of *interpretability* of AI systems. It is argued that "[p]roviding some method for effectively communicating an agent's beliefs, desires, and plans to the people around it is critical for ensuring that artificial agents act ethically".[1171] A possible solution could be if the agent were to "*explain its reasoning* by describing its predictions of the consequences and how it thinks those consequences are valued".[1172] Such a statement is relevant for this research since interpretability, i.e., the ability for AI systems to communicate their (moral) decisions, is considered by some as a *fundamental condition for criminal liability.*

As a matter of fact, Hu argues that interpretability is relevant for criminal law purposes, as only by understanding the *reasons* behind a certain act it is possible to ascertain whether the robot committed the action with the required *mental state*. In other words, Hu links *interpretability* to *mens rea*. When doing so, she imagines the following commission-by-omission scenario:[1173] let us imagine that a robot is passing by, when it hears a child crying for help because she is drowning. We can imagine that the robot – similar to Abel's Burning Room ethical dilemma discussed above[1174] – will make "a series of calculations: If it stops to pull the child out of the water, there is a ten percent chance that it might fall into the water, destroying itself. If it does not, there is a ninety percent chance that the child will die".[1175] I agree with Hu that, in scenarios like these ones, we might be able to hold AI systems to a higher moral standard than humans.[1176] Not only one could argue that it would be morally wrong for the robot not to risk its "life" to save the drowning child, but we could even imagine imposing a general legal duty on AI systems to save human beings, no matter the costs, and hold them *accountable* if they fail to do so. Nevertheless, this would entail that we,

---

[1169] Hörnle, 2014, p. 693.
[1170] See Ch. 4.2.
[1171] Abel, MacGlashan & Littman, 2016, p. 7
[1172] Abel, MacGlashan & Littman, 2016, p. 7 [emphasis added].
[1173] See Example D at Ch. 1.6.
[1174] See Ch. 5.3.2.2.
[1175] Hu, 2019, p. 500.
[1176] Hu, 2019, p., 500.

as a society, would consciously decide to assign a lower value to the "life" of an AI system than to the one of human beings.

Hu takes this reasoning even further, since she claims that we could hold AI systems *criminally* liable for not acting like heroes.[1177] I believe that even though it might be feasible to do so, it would be uncalled for, since criminal law is not the only tool of reprimand available in a society. I will expand on this further on in this Chapter.

How would this impact the creation of AMAs? It would call for a different "ethical training" from the one that we – as human beings – receive when growing up.[1178] Surely, teaching a robot how to make the decision of not (always) being a hero, which would be akin to the ethical decision that we would take as humans, while at the same time punishing it when it decides not to be one, would be counterintuitive.[1179] In this I find the greatest limitation of Hu's theory, which is based on a futuristic, if not even dystopian, scenario. As it will be discussed in the following paragraph, it would be different if we were faced with robots who could learn ethical values autonomously.

8.4.1.2 The (Ir)Relevance of Motives

Criminal codes are not a codification of moral wrongs. As it is often taught in the first class of a criminal law course, not everything that is immoral is illegal and not everything that is illegal is immoral. Yet, criminal culpability is "grounded in moral blameworthiness".[1180] It "seeks to refine and institutionalize our intuitive blameworthiness judgments"[1181] and can be understood as a "stripped-down, coarse grained analog of blameworthiness".[1182]

[1177] Hu, 2019, p. 500.

[1178] Our "Innate Moral Core". See J. K. Hamlin, "Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core", *Current Directions in Psychological Science* Vol. 22, Iss. 3, 2013, pp. 186-193.

[1179] E.g., if we introduced an offense in the criminal code for robots targeting the failure to save a human being. See also Consulich, who argues: "*L'AI non è infatti suscettibile alla minaccia di una qualsiasi forma di sanzione, tanto quanto sia in prima persona il destinatario della sanzione quanto allorché lo sia un suo 'simile'. Occorrerebbe programmare forme che contengano l'istruzione di essere recettivi rispetto a tali input, ma si tratterebbe di una soluzione al limite dell'autopoiesi: costruire una sensibilità al precetto giuridico per poi creare una responsabilità, civile o penale che sia*". F. Consulich, "Le prospettive di *accountability* penale nel contrasto alle intelligenze artificiali devianti", *Diritto e Procedura Penale*, Vol. 3, 2022, p. 1022.

[1180] A. Sarch, "Should Criminal Law Mirror Moral Blameworthiness or Criminal Culpability? A Reply to Husak", Law and Philosophy, Vol. 41, 2022, p. 309.

[1181] Ibid.

[1182] Ibid.

Let us now focus for a moment on what we have, i.e., (human) criminal codes. Modern criminal legal systems seem to be slightly schizophrenic: on the one side, they demand that criminal agents have the *general capacity to be responsive to the law*, i.e., the ability to *motivate themselves* in accordance with the law; on the other side, they do not demand a *specific motive* to commit a *specific offense*. In a nutshell, this is the so-called "irrelevance-of-motive maxim".[1183]

In the example of the robot "hero" and the drowning child, described above, Hu argues that the *mens rea* requirement would be fulfilled by the robot: (a) if the robot knew that the drowning child was actually a human child; (b) if the robot knew that the drowning child was in imminent danger; (c) if the robot was able to carry out actions that help reduce the danger; (d) if the robot *concluded* not to take any of the actions identified in (c).[1184] Whether it decided not to save the child because it preferred not to risk its integrity, or whether it deemed more important a different task, such as delivering a package, would not be relevant when establishing its liability (i.e., it would be an irrelevant *motive*). Hence, in such a situation the robot would be liable – provided that a norm establishing said liability is in place. Contrarily, the robot's liability could be questioned if it erroneously believed to be in the process of saving another child (who is instead already dead), or if it two kids were drowning and it had to choose between them.

Hu believes that criminal liability could be imposed only if the robot were able to communicate the moral *reasons* which led it to make a certain decision, or else "it would be difficult, if not impossible, to determine whether a wrongful action has taken place".[1185] I maintain that this communication could only lead to establishing a *motive* for the robot's action, not its *mens rea*.

Let us take a step back. When examining whether the *mens rea* element of an offense has been fulfilled by one's conduct, judges make use of certain objective elements that work as clues of the agent's *animus necandi* (e.g., the murder weapon and the number of lethal blows). No relevance is given to the fact that the offender might have murdered her neighbor because she simply didn't like her. Motives can have relevance as exculpatory factors only whenever they are indicators of a lack of criminal capacity (i.e., when they can be hints of a

---

[1183] See *ex multis*, M. T. Rosenberg, "The Continued Relevance of the Irrelevance-of-Motive-Maxim", *Duke Law Journal* , Vol. 57, No. 4, 2008, pp. 1143-1177.
[1184] Hu, 2019, p. 511.
[1185] Hu, 2019, p. 512.

legally relevant insanity), provided that they are supported by evidence on the mental health status of the offender.[1186] This could be the case of the robot who – in the scenario above– believed that it was in the process of saving another drowning child, when instead the child was already dead, due to a defect in its system caused by a virus.

The rationale behind the irrelevance-of-motive maxim lies in the differentiation between *motive* and *intent*: the first is a "ground or reason for action",[1187] not a "driving force that compels a person's actions";[1188] the second is "the objective effect which the law-breaker contrives to produce on others by his act or omission".[1189] Motives are "the psychic cause, the stimulus, the spring, the impulse, the feeling, the instinct, that prompted, moved, induced, the subject to act (or to omit)".[1190] They are the factors that triggered the will and that eventually determined the individual to commit the crime.[1191]

Accordingly, "motive answers the question why [*a person acted as he or she did*], neither in terms of causation nor in those of a further ulterior objective, but in terms that give a reason which is the subject of an *ethical appraisal*".[1192] Instead, intent entails the *deliberate* and *conscious* effort to engage in *behavior that goes against the law*: "[t]o understand why a defendant took the action he did, the explanation proceeds in terms of intended consequence with each intention explaining the preceding intention".[1193] The evaluation of motive is essential to "judgments of moral culpability"[1194] and is not directly relevant for judgments of criminal liability. The two elements are, naturally, connected, yet they must be kept separate. As a matter of fact, intention relates to the moment and stage of the decision making process *between different reasons for acting* differently.[1195]

As stated above, one could perhaps identify *motives* in the commission of crimes by (future) AMAs. Consequently, harm caused by AMAs could be deemed wrong from a *moral* perspective – and be subject to the social consequences that can follow the breach of a moral

---

[1186] Motives can also have relevance as aggravating or mitigating factors in the determination of a sentence.
[1187] J. Hall, *General Principles of Criminal Law*, 2nd Ed., The Bobbs Merrill Company, 1960, p. 88.
[1188] Ivi, p. 89, n. 77.
[1189] Ibid.
[1190] P. Veneziani, *Motivi e colpevolezza*, Giappichelli, 2000, p.3.
[1191] Ibid.
[1192] Hall, 1960, p. 93 [emphasis added].
[1193] Rosenberg, 2008, p. 1157.
[1194] Ivi, p. 1166, note 128.
[1195] A. Grandinetti, "Motivi e movente", *Il Penalista*, Bussola. Available at: https://ilpenalista.it/bussola/motivi-e-movente.

rule. The previous paragraph was concluded with the ideas of robots learning ethical values autonomously: until systems like this are developed – and it is not sure that they will ever be – pinning blame on these systems would probably be superfluous. [1196] Here I take the reasoning even further and argue that, since AMAs are not able to experience the negative feelings which follow moral responsibility, such as guilt, disapproval, or reproach, it seems difficult to qualify them as true moral agents for now. As it will be argued in Paragraph 8.5, empirical research might prove just the opposite.

Moving forward, I find that there is one major obstacle to affirm that AI systems can display *intent* and, more generally, to arguing that AI systems can possess criminal capacity: AI systems lack *agency* (*and the feeling of it*), which is the probably the most human characteristic about humans. As it was effectively argued,

> by far, a characteristic feature of natural beings that has been largely ignored by engineers, is the *sense of "self"* that pervades across all cells of the organism. Cells have a sharp notion of "citizenship" to the being that make them act with vigilance against "foreign" cells that infect the organism. Even though each agent in the being is acting autonomously, there is also a sense of "oneness" or "belongingness" to the being, that pervades across all the agents. When the organism strives to survive, it is the pervasive sense of self that is sought to be maintained and preserved and not for instance, any particular cell. It is also the pervasive sense of self, that is sought to be protected against attacks in the form of infections, by the immune system.[1197]

[1196] Kalliokoski and Hallamaa make the following example when denying moral agency to AI systems. Let us think of an AI system whose purpose is to assign doctor appointments to patients. When doing so, the AI system will have to assign different grades of priority to the patients. To make this decision, the AI system should include parameters such as "the patients' age, social status, and indicators concerning vulnerability". Conceivably, these parameters "would improve the monitoring of appointments in medical care in terms of fairness—which is a moral characteristic—but it would not make the AI system morally responsible for placing the patients in a preferential order" since those who "feel they have unjustly been left waiting in the patient queue would, rightly, blame those who designed the code for the monitoring system rather than the AI application itself". T. Kalliokoski , & J. Hallamaa, "How AI Systems Challenge the Conditions of Moral Agency?", in M. Rauterberg (Ed.), *Culture and Computing: 8th International Conference, C &C 2020, Held as Part of the 22nd HCI International Conference*, Springer, 2020, p. 8.

[1197] Srinivasa & Deshmukh, 2021, p. 3 [emphasis added].

Accordingly, only an agent who has an "experience of authorship",[1198] i.e., that can experience a "constellation of feelings",[1199] such as being able to "perceive herself as the cause of her own actions tracking the linkage between a voluntary bodily movement and its effect",[1200] can be deemed as *responsive* to the command expressed in the law, specifically by criminal offenses.[1201] In human beings, agency is not a "binary property"[1202]: it emerges gradually during "the physical, social, and psychological development of a child".[1203] In a humancentric legal system, agency establishes that link between someone's action and her criminal liability. As it was maintained by R.A. Duff, "intentional agency" is a central paradigm of "*responsible agency*".[1204] In a criminal legal perspective, one could argue that criminal liability can be construed as being addressed to those who are "rationally responsive" to criminal law's remands.[1205] It follows that only those who *question the validity of a norm* can be considered as "true" authors of crime.[1206]

As it was claimed, intent and, more generally, criminal capacity, denotes being receptive to the *prescriptive* contents of a criminal norm, i.e., its persuasive and coercive components. In this sense, scholars argue that not only does the criminal norm "orient" an individual's behavior through the threat of the sanction – which will be inflicted if she harms a protected legal goods, but also that it is internalized by the individual and becomes part of her decision-making process.[1207] In other words, criminal law, either because of the fear of being punished, or other cognitive mechanisms, delivers a message which *impacts* individuals and *guides* their behavior.[1208] This is possible because human beings are *susceptible* to such message. Current AI systems, as this research showed, are not.

[1198] Bonicalzi & Haggard, 2019, p. 113.
[1199] Ibid.
[1200] Ibid.
[1201] Ibid.
[1202] Kalliokoski & Hallamaa, 2020, p. 4.
[1203] Ibid.
[1204] R.A. Duff, *Intention, Agency and Criminal Liability: Philosophy of Action and the Criminal Law*, Blackwell, 1990, p. 101 [emphasis added].
[1205] V. Chao, *Action and Agency in the Criminal Law, Legal Theory*, Cambridge University Press, 2009, p. 23.
[1206] Simmler & Markwalder, 2023, p.30.
[1207] C. Colucci, *Tra ottimizzazione della funzione comando e prospettive di un suo superamento: i nuovi scenari della normatività penale*, PhD Dissertation, University of Florence Repository, 2022, pp. 4-5. Available at: https://flore.unifi.it/handle/2158/1273584.
[1208] "*L'agire umano, infatti, è psichicamente (e non meccanicamente) causato, e cioè determinato mediante motivazioni. Nella struttura normativa del comando tale motivazione è rappresentata dalla presenza della sanzione, e cioè dalla minaccia della*

Only if AI systems could be qualified as true "*addressees of normative expectations*"[1209] we would be able to hold them criminally liable, since it would imply that the machine made a mistake, i.e., it committed a crime, even though *it could have acted differently*.[1210] And it is only in this light that its punishment would incorporate an expectation of a "a different norm-conforming behaviour"[1211] in the future.

8.4.2 Retributivists at Heart?

In the previous paragraph I argued that being an addressee of criminal law entails rationality, meant as the general capacity of being susceptible to legal commands and of acting upon them freely (or, better, with a *perceived sense of freedom*). Rationality, in an anthropocentric legal system, is meant as the ability for the agent to motivate herself to behave in a way that conforms with the law – through the anticipation of the consequences of her act/omission (*actus reus*). This is the quintessence of criminal capacity, which is a quality that must pre-exist in criminal agents.

As it was mentioned in Chapters 4 and 5, criminal capacity is strictly connected to *purposes* of criminal punishment. As such, this study necessarily touched upon different theories of punishment. Let us now focus specifically on retributivism and on the "retribution gap"[1212] issue, a matter that was brought to attention in the preliminary conclusions of Chapter 6.

---

sua applicazione nel caso in cui il precetto sia disatteso. Tale contromotivo […] in una concezione della norma che si inspiri alla centralità del momento sanzionatorio, rappresenta l'unico perno psicologico della norma stessa che, per mezzo dell'intimidazione, persegue lo scopo di impedire la lesione di beni giuridici da parte dei destinatari della norma stessa. […] Se questo è dunque il meccanismo psicologico per mezzo del quale funziona il comando, è pur necessario mettere in evidenza come i meccanismi di orientamento del comportamento – anche di carattere normativo – siano molto più vari". Ivi, pp. 75-76. The fact that human beings are *susceptible* to such message, though, does not mean that one must give in to deterministic theories, i.e., consider humans as "pre-determined" beings: "*è necessario […] mantenere ferma l'idea che la prospettazione di un comando, al quale è legata la minaccia di una sanzione, attivi nel destinatario un processo cognitivo pienamente consapevole, senza cedere alla tentazione di ricondurre anche la paura di una conseguenza negativa all'insieme di istinti naturali che, non pienamente controllabili dall'uomo, appartengono al mondo dell'essere*". Ivi, p. 78.

[1209] Colucci, 2022, p. 78 [emphasis added].

[1210] Ibid.

[1211] Ibid.

[1212] Danaher, 2016.

Retributivism is, in essence, "the idea that what justifies criminal punishment is that it is deserved for past criminal wrongdoing".[1213] It entails answering the questions of: "*what* criminal wrongdoers deserve to suffer; *why* they deserve to suffer it (how it is that crime makes such suffering appropriate); and *why* it should be the business of the state to create and maintain an institution whose purpose is to impose that deserved suffering" – the so-called "just deserts" doctrine.[1214] Hence, criminal punishment must be "burdensome, or in some sense painful":[1215] this does not entail that punishment should be meant *only* as the imposition of a burden or the delivery of pain, since it also serves a *communicative* function. As such, "it communicates directly to the offender, but also to all citizens, the censure that the crime deserves"[1216] and for that reason "it is justified as a response to the wrong for which it is imposed and must be appropriate in its character and severity to that wrong".[1217]

The question which we ask ourselves here then is: in order to be responsive to the criminal sanction, does the agent need to be able to feel pain? Does rationality mean responsiveness to the threat of the infliction of harm? As a matter of fact, it strikes as odd that – when dealing with such an innovative topic as liability of AI systems – we seem to reduce ourselves to traditional retributivists arguments. The "new vs. old" clash is glaring. Indeed, retribution – as compared to other "instrumental goals of punishment"[1218] – is backward-looking: it is how the offender repays society for the harm she caused. Deterrence, on the other hand, is forward-looking: it deals with the "future benefits of reducing the likelihood of crime".[1219]

If one focuses on the "communicative" function of criminal punishment, it is doubtful whether punishing AI systems would create benefits for society as a *whole*. As a matter of fact, in order to call someone (or something) to publicly account for wrongdoing it must be "a fellow member of a relevant normative community … a responsible agent who can be

---

[1213] R.A. Duff, "Responsibility, Restoration, and Retribution", in M. Torny (Ed.), *Retributivism Has a Past: Has it a Future,* Oxford University Press, 2011, p. 63.
[1214] Ivi, p. 65.
[1215] Duff, 2011, p. 66.
[1216] Ivi, p. 78.
[1217] Ibid.
[1218] M. Gerber & J. Jackson, Jonathan, "Retribution as revenge and retribution as just deserts", *Social Justice Research*, Vol. 26, No. 1, 2013, p. 63.
[1219] Ibid.

expected to answer for what he has done".[1220] It was contended that AI systems are not part of our moral nor normative community and, as such, *their punishment might not fix what they broke*. Furthermore, if they were considered as agents part of one's moral community, they would also have the right to be safeguarded from unjust criminal punishment – unless one would want to attribute to them a status similar to those of slaves.

It is plausible, though, that the punishment of AI systems would have *indirect deterrent effects* on human subjects. It could influence certain human agents into doing everything they can to avoid that an AI system, that they produce or use, commits a crime, since their punishment (be it through reprogramming or destruction) would damage them economically (and perhaps, in certain cases, also personally).[1221] This topic will be addressed further in the following section.

Imaginably, one could argue in favor of a change of paradigm from conceiving criminal sanctions as suffering to conceiving them as a tool to influence behavior. Actually, this change of paradigm is already palpable in our (risk) society, which is characterized by the proliferation of endangerment offenses, and is also one of the pillars grounding CCL. At this point, one must ask herself: if we strip criminal law of punishment, can we still call it criminal law? In the eyes of a criminal lawyer, the bond between crime and punishment appears as unbreakable. Finally, holding AI systems criminally accountable calls for a thorough examination of conscience: are we all – deep down – retributivists?

8.4.3 Re-Evaluating the Comparison Between AI and Corporations

In Ch. 6, it was argued that AI systems make for very (hypothetical) rational agents, according to an economical concept of rationality as utility-maximization. This is akin to what has been argued about corporations, who, "while lacking feeling and emotions nevertheless possess capacities for intelligent agency".[1222] As a matter of fact, "lack of emotions in this regard may

---

[1220] Duff, 2011, p. 6.

[1221] "The idea is that if users/deployers were fined (for amounts exceeding compensation) for crimes committed by AI systems, this should induce them to prevent such crimes. In particular, they should be induced to provide AI systems with the motivation to act in such a way as to prevent sanctions against their users. This might be obtained—when deterrence through sanctions is the most appropriate way to influence the behaviour of AI systems—either by making such systems internalise in their utility function the disutility resulting from sanctions against their users, or by providing adequate private sanctions". Lagioia & Sartor, 2020, p. 29.

[1222] Keiler, 2013, p.51.

accordingly foster rather than hamper rational choice"[1223] and may make a good case for punishing AI systems to achieve deterrence.[1224] In this perspective, rationality makes corporations (and AI systems) ideal candidates for a criminal sanction, since they are not subject to irrational impulses nor to biological reactions such as intoxication or brain damage.[1225]

When discussing liability of AI systems, one starts with the same premises as with CCL: "human concepts", such as *actus reus* and *mens rea,* are "not readily applicable in the realm of corporate [*nor algorithmic*] wrongdoing".[1226] The attribution of criminal liability to corporations, especially in those systems that adopt an "organic" model for CCL, implies a change of archetype. Undeniably, "the paradigm of moral blameworthiness ought to be applied differently to corporations".[1227] Assuming that one can identify in corporations certain capacities, such as rationality and autonomy, which are premises of criminal liability, they will necessarily "acquire a different connotation … as opposed to individual criminal responsibility".[1228]

According to some authors, CCL "demonstrates a certain degree of flexibility shown by criminal law when criminal policy demands so. A flexibility that can be used provided that certain dogmatic premises are met, to justify the punishment of AI entities".[1229] CCL, therefore, can be seen as an example of the irrelevance of distinguishing "right from wrong" for the purposes of criminal liability.[1230] As such, it can provide insight on how to overcome the challenges that were highlighted throughout this study with regards to the capacity of AI systems to make "ethical" decisions. Possibly, CCL supports the idea that moral

---

[1223] Ibid.

[1224] "Deterrence theory is a powerful and flexible approach for designing criminal justice policies. At heart, it is an economic theory. It views both would-be and actual criminals as rational actors who are trying to maximize their utility. Deterrence theory is a powerful and flexible approach for designing criminal justice policies. At heart, it is an economic theory. It views both would-be and actual criminals as rational actors who are trying to maximize their utility […] The task for deterrence theorists is to hit the sweet spot where sanctions are just high enough to prevent these crimes". M. Diamantis, "Clockwork Corporations", *Iowa Law Review*, Vol. 1032, 2020, pp. 518-520.

[1225] Diamantis, "Clockwork Corporations", 2020, p. 519.

[1226] Keiler, 2013, p. 51.

[1227] Keiler, 2013, p. 53.

[1228] Ibid.

[1229] Freitas, Andrade & Novais, 2014, p. 11.

[1230] Such an ability is deemed irrelevant also for humans in certain legal systems. See the recent US Supreme Court judgment Kahler v. Kansas, discussed in Ch. 5.2.

blameworthiness "is not the criminal justice system's sole concern, and perhaps not even its primary concern".[1231] The focal goals of CCL are, on the one hand, the prevention of harm to society and, on the other hand, the deterrence of conducts which could lead to said harm, "regardless of whether the conduct in question can be characterized as immoral.[1232] So what makes corporations different from AI systems? And what would be the political reasons behind creating AI criminal liability?

To begin with, applying the concept of *respondeat superior* to AI systems seems to be tricky. AI systems are not made of humans who follow a fixed internal hierarchal structure. Hence, at first glance, there is no identifiable *master* who would be liable for the actions and the *mens rea* of its subordinates.

Moreover, as it was argued by John Coffee,[1233] punishing corporations ultimately entails influencing the decision-making processes of corporations "so that they do not do the same thing again.".[1234] Admittedly, the "normative" behavior of corporations can be controlled through criminal sanctions, even though they have no "body to kick",[1235] *as long as there are (identifiable) humans behind them.*[1236] Punishing corporations might prove difficult since they are "complicated distributed systems".[1237] It follows that in order to make punishment effective, one must know "the *structure* of the organization" and "how decisions are made on different levels of the organization"[1238] so that she can "apply incentives and penalties that will influence the values of those decisions and their utility functions there".[1239] Nevertheless,

---

1231 Gobert & Punch, 2003, p. 46.

1232 Ibid. See also Regina A. Robson, who claims that "virtual elimination of retribution as an acknowledged goal of [corporate-]criminal sanctioning," with only deterrence left standing to explain why the State should hold corporations criminally responsible". R. A. Robson, "Crime and Punishment: Rehabilitating Retribution as a Justification for Organizational Criminal Liability", *Am. Bus. L.J.*, Vol. 47, 2010, p. 121.

1233 J. C. Coffee Jr, " 'No Soul to Damn: No Body to Kick': An Unscandalized Inquiry into the Problem of Corporate Punishment", *Michigan Law Review*, Vol. 79, No.3, 1981, pp. 386-459.

1234 Asaro, 2014, p. 290.

1235 Quote by Edward, First Baron Thurlow 1731-1806 reported in John C Coffee Jr, "No Soul to Damn: No Body to Kick": An Unscandalized Inquiry into the Problem of Corporate Punishment' (1981) 79 Michigan Law Review, 386-458.

1236 A. Peikert (Dipl.-Jur.), A. Reinelt (Dipl.-Jur.) & J. Witt (Dipl.-Jur.), "Diskussionsbeiträge der 38. Tagung der deutschsprachigen Strafrechtslehrerinnen und Strafrechtslehrer 2019 in Hannover", *ZSTW*, Vol. 131, No. 4, 2019, p. 1133.

1237 Asaro, 2014, pp. 290-291.

1238 Ivi, p. 291 [emphasis added].

1239 Asaro, 2014, p. 291.

corporations seem to have a "collective soul" to damn, as it is synthetized in the doctrine of corporate legal culture. It is much harder to transpose the same concepts to AI systems even though, sometimes, such in the case of robots, they do have a body to kick.

In other words, by punishing corporations we can influence their behaviors because those within the organization "get the message". When it comes to AI systems, detecting a collective soul and a collective culture is problematic. There is not a clear internal and hierarchical structure behind it. However, the lack of such a structure might only be apparent. Indeed, the development of an AI system can involve plenty of human subjects accompanied by a fragmentation of roles and responsibilities. The combination of their personal attitudes could lead to the creation of a sort of *temporary collective agent*, who would have its own attitude, separate from those of its members, similarly to what happens with conspiracies to commit crimes.

Thus, I argue that this alone is not enough to introduce criminal liability of AI systems. The real issue is that AI systems cannot possess funds and they do not possess a "profit motive,"[1240] therefore punishing them via fines would be useless. As it was argued, "[a]t least with corporations, they exist essentially to make money, so taking money away from them actually hurts them, because that is their fundamental reason of existing".[1241] It follows that AI systems could only be punished via more drastic measures, such as incapacitation. Nevertheless, these measures are likely to only impact the humans-behind-the-machines indirectly, as AI systems "do not care about their physical presence or state".[1242] As a consequence, putting them in prison is not really "going to change their minds about anything or effectively punish them if they do not care about their bodies".[1243] Indeed, "*if they do not really care about anything, how do you punish them at all?*".[1244]

Perhaps, the indirect effects of punishing AI systems could be the criminal policy motive to introduce such a liability framework. Once again, it is a matter of assessing what would be the ultimate goal of criminalizing a specific conduct. Criminalizing *directly* AI systems would *indirectly* penalize those who benefit or profit from the AI system, while at the same time it would minimize "the need to prove that the harm was attributable to specific natural persons

[1240] Ibid.
[1241] Ibid.
[1242] Ibid.
[1243] Ibid.
[1244] Ibid. [emphasis added].

or corporations".[1245] Thus, in the case of AI systems, the plethora of subjects which would suffer the consequences of a criminal AI systems requires better definition. It is for this reason that Diamantis' theory proves successful: it acknowledges that most AI systems are produced and deployed by organized corporations and, by relying on an already consolidated legal framework, it provides a solution to address AI harm almost effortlessly. In this perspective, punishing AI systems could for example lead to an increase in the duty of care of CEOs and other high-ranked employees. Nevertheless, , if the purpose is to deter those who are responsible for the AI system, I agree with those who claim that there would be equally efficient alternative mechanisms.[1246] I will come back to this statement in the following paragraphs.

8.4.4 Epicenters of Liability: the Human Culprit

This research was centered around criminal liability *of* AI systems. Yet, it also touched upon criminal liability *of* humans *for* AI systems. The topic was addressed from a theoretical perspective at Chapter 6.2.3., which discussed whether humans-behind-the-machine could be deemed responsible for AI harm based on a negligence *mens rea* standard; and at Chapter 7, which instead scrutinized the topic from a quasi-*de iure condito* perspective.[1247] I will now draw some conclusive remarks on the matter and outline avenues for future investigation.

Automation is safer and superior to human beings, until it's not. Surely, the interaction between AI and human beings leads to an enhanced risk of moral disengagement by the human-behind-the-machine.[1248] Moral disengagement is an expression coined by psychologist Albert Bandura to refer to a set of techniques by which human beings "selectively disengage moral self-sanctions" [1249] when doing something "wrong".[1250] As a

---

[1245] SAL Report, p. 38, para. 4.47.

[1246] Ibid.

[1247] The examples analyzed at Chapter 7 are not all codified law.

[1248] See S. Massi, "Affidamento sull'intelligenza artificiale e 'disimpegno morale' nella definizione dei presupposti della responsabilità penale", in R. Giordano, *Il diritto nell'era digitale. Persona, Mercato, Amministrazione*, *Giustizia*, Giuffrè, 2022, pp. 665-679.

[1249] A. Bandura, "Moral disengagement in the perpetration of inhumanities", *Personality and Social Psychology Review*, Vol. 3, 1999, pp. 193-209.

[1250] Bandura mentions eight mechanisms of moral disengagement: moral justification; exonerative comparison; euphemistic labeling; minimizing, ignoring or misconstruing the consequences; dehumanization; attribution of blame.

matter of fact, AI systems might act in ways that are obscure to their users or creators, and this compels them to just "trust the (technological) process", weakening their sense of agency. This reasoning could be taken even further to claim that human agents can convince themselves that what they're doing – when dealing with a machine – is not only morally acceptable, but also legally acceptable. The complex relationship between one's sense of agency and its responsibility was already addressed before. How does moral disengagement, then, impact the attribution of negligence upon the human-behind-the-machine?

The case studies reported in Chapter 7 represent the first instances of regulators attempting at striking a balance on when criminal law should "track" morality and when it should not in the field of criminal liability *for* AI systems. They concern exclusively liability of *humans* for actions of *non-humans*. Their collection and comparison is one of the elements which make the present research original.

Criminal law is a matter of design: criminal legal systems are the results of a set of practical decisions ("institutional design considerations")[1251] which can include "bright line rules",[1252]such as the ones suggested by the UK Law Commissions, and "generally applicable resolution of contested normative questions".[1253] The majority of the case studies examined in Chapter 7 rely almost entirely on the human's ability to anticipate and manage risks, i.e., on her negligence. As it was argued, the application of a such a standard of *mens rea* could give rise to "negligence failures".[1254] This term describes "situations in which the classical building blocks of negligence, i.e., risk taking, foreseeability, and awareness, struggle to identify a liable human being to whom we can attribute AI-caused harm".[1255]

Negligence, generally speaking, requires that the agent foresaw and accepted the risk that an unlawful consequence would arise from her conduct. Foresight can be meant as a *generic* knowledge that the activity could lead to some harm, or as the knowledge that the activity could actualize a *specific* risk, i.e., a specific harmful consequence.

[1251] Sarch, 2022, p. 310.
[1252] Ibid.
[1253] Sarch, 2022, pp. 310-311.
[1254] Negligence failures represent a direct development of Abbott and Sarch's irreducibility challenge. Giannini & Kwik, 2023, p. 3.
[1255] Ibid.

We asked ourselves before if humans will always be able to effectively supervise a system that was created to overcome them and make up for their shortcomings.[1256] Now we can answer that probably they will not. Nevertheless, they represent, for now, the only *liable* entities which are most proximate to the AI system. As such, I believe that they could be deemed liable for the AI's misbehavior, provided that they negligently disregarded precise best practices and technical standards (which still have to come to life) and that this led to the commission of a *grave* offense (e.g., negligent homicide).

Moreover, the standard of care of the model agent should be modelled according to single area of application of the AI systems (e.g., toys, transportation, healthcare) and to the role covered by the human agent in question (e.g., trainer, tester, data collector). On the contrary, I maintain that holding human beings criminally responsible for the sole fact that they created, used, or produced an unpredictable machine would be excessive. In these cases, the best solution is to be found in other areas of the law, such as torts.

As it was suggested to in Para. 4.1., one could take inspiration from the field aviation and, specifically, from the liability of air traffic controllers.[1257] Air traffic controllers have a criminally relevant duty of care: they are obliged to prevent collisions between aircrafts and, more generally, to avoid air disasters. Thus, they are expected to be highly skilled and trained subjects. They operate in a field where human and technological errors blur. In this sense, one would need to identify with precision who could cover a similar position in the chain of creation and deployment of an AI system, taking into consideration that in most situations end-users will be simply "average citizens". Indeed, the attribution of liability for AI-crimes could benefit from a field like the one of major disasters: their complexity, both with regards to the web of causation and to the high level of automation, requires a multidisciplinary approach.

Moral disengagement, lack of foresight, lack of meaningful control: all these conditions (and many others), taken together or singularly, can reduce heavily – or even eliminate – the knowledge that an AI system could behave criminally, hence leading to a negligence failure. The UK, Singaporean and French examples, when "drawing a bright line" amid "risk zones of legality" and "zones of illegality", necessarily strike a balance between risks of scapegoating

---

[1256] Above at Paras. 3.4.1 and at 8.4.4.

[1257] E. Greco, *Profili di responsabilità penale del controllore del traffico aereo. Gestione del rischio e imputazione dell'evento per colpa nei sistemi a interazione complessa*, Giappichelli, 2021.

humans-behind-the-machine and the need of legislators to regulate new technologies. Only (judicial) practice will tell whether the balance was struck fairly.

8.5 GENERAL CONCLUSIONS. THE FUTURE- AND BACKWARD-FACING QUESTION

The time has come to focus on the main RQ guiding this research: to what extent is a theoretical framework of criminal liability for non-human agents *needed* and *feasible*?

The RQ and the title of this research clearly reveal that the starting point of this research was the matter of *direct* liability of AI agents. Approaching such a topic demanded a certain level of *flexibility*: as it was shown in Chapter 3, several criminal legal scholars refuse to even consider such a matter, discarding it as absurd. Apropos, the choice of a RQ formulated in terms of "to what extent" allows for this plasticity, since it grants the researcher the freedom to explore a certain field without incurring in strong "yes-or-no" alternatives.

After a close review, it is suggested here that the two prongs of the research question are switched, i.e., that the RQ is reformulated as follows: to what extent is a theoretical framework of criminal liability for non-human agents *feasible* and *needed*? Perhaps, the original version of the RQ was more compatible with criminalization theories. Nevertheless, the development of AI systems touches upon conducts that have already been reckoned as worthy of criminalization. Hence it appears more logical to first discuss whether it would be practicable to adapt what is already in place, i.e., traditional criminal norms, and then address the question of whether to attribute liability to AI systems or not, since the latter is essentially a question of criminal policy. I assume that this type of reasoning would be also followed by policymakers, who would have to justify the necessity of introducing new legal concepts on its feasibility. Indeed, one could have the greatest idea of all times, but if it is not feasible, is it really a great invention?

Let us focus, then, on what constitutes (now) the first prong of the RQ (to what extent is a theoretical framework of criminal liability for non-human agents *feasible*?). Studying feasibility entails analyzing critical aspects of a proposition in order to determine whether it would succeed.[1258] In the case of this research, it called for an inquiry into whether AI systems

[1258] Investopedia, "Feasibility Study", 4 November 2022. Available at: https://www.investopedia.com/terms/f/feasibility-study.asp.

would be *eligible addressees* of criminal law. To do so, I broke criminal liability into three building blocks (criminal capacity, *actus reus, mens rea*) and tested whether each one of them could be fulfilled by AI.

The succinct answer to the first prong of the RQ is the following: criminal liability of AI agents is *feasible* depending on how much of our anthropomorphic notions we are willing to give up. Specifically, holding AI systems responsible would entail either being able to create machines that are *susceptible* to the commands of a legal norm, or abandoning the concept of rationality (meant as the ability to be responsive to the law) altogether. Indeed, I advanced above that the *comprehension* of an offense's command is an essential prerequisite for establishing criminal liability and, as such, it represents the main obstacle to attributing liability to AI systems directly. In fact, criminal liability can be imputed only to individuals who are capable of understanding and following the *demands* of criminal law. This means that in order to affirm that a machine committed a crime, we must be certain that it made a deliberate choice *not* to conform to normative expectations, even though it had the capability to do so. As of now, it is not possible to attribute this capacity to modern AI systems. They may consequently be held criminally liable in the future if this condition were to be satisfied.[1259]

In other words, AI just doesn't get it. It does not fit in the "insufficient regard picture"[1260] painted by a criminal offense. AI systems cannot be criminally liable as long as a criminally culpable act (in contrast to a morally blameworthy one)[1261] is understood according "to the degree it's based on a valuation of reasons that diverges from the correct weights the law ideally would say should be ascribed to the reasons bearing on whether to do the action".[1262] Besides, it would be *feasible* to establish criminal liability *of* AI agents only if we were to abandon the concept of retributivism as a goal of punishment in the foreseeable future. That is because not only AI systems do not comprehend the essence of the criminal norm, but also they do not comprehend its sanction.

[1259] In this sense, I agree with Simmler & Marwkalder, 2019. p. 29.

[1260] "For the interests, rights and values that are properly protected by criminal law", Sarch, 2022, pp. 309-310.

[1261] An act is morally blameworthy "to the extent it manifests (in the way morality requires) insufficient regard for the interests, rights, and values morality deems relevant". Ivi, p. 309.

[1262] Sarch, 2022, p. 310.

Let us now focus on the second prong of the RQ (to what extent is a theoretical framework of criminal liability for non-human agents *needed*?). My answer is that criminal liability of AI agents is not *needed* as long as AI systems are not treated as members of our community and, as such, as subject to the same rules as us. The change in perception, i.e., the anthropomorphization of AI Systems, could happen only the moment "true" AMAs are developed. As argued above, being an AMA would suffice to pin moral blameworthiness on AI systems, but in order to be considered criminally culpable they should also be "legal" agents. And, in other to be *legal* agents, AI systems would have to be responsive to criminal norms. In other words, the questions of whether criminal liability of AI systems is feasible and needed remain intertwined.

If we were to ascertain that criminal liability of AI agents is needed, we would also have to accept that they would have the right to have a say regarding their punishment, unless one is willing to accept the creation of a new generation of "robotic" slaves. Notably, this research focused "only" on AI systems and their legal personality as "duty bearers",[1263] i.e., on whether they could be subject to sanctions. The result of this study could be of use for future research on the question of whether AI systems could also be "right-bearers",[1264]specifically, in the realm of criminal law, of whether they could be deemed victim of crimes.

Surely, the creation of a new sub-system of rules for AI systems should be avoided. Modern criminal law is democratic: it applies to every citizen, according to the principle of the rule of law. Apropos, "[i]t's important that the public perceive the criminal law to be treating like cases alike, as this bolsters the institution's perceived legitimacy and engenders trust and obedience to its standards".[1265]

Furthermore, one must also take into account what would be the goal of criminalizing AI systems, also in comparisons with alternatives to criminal law. To put it another way, we need to ask ourselves whether criminalizing AI systems would be *useful*. As a matter of fact, according to the principle of subsidiarity, before criminalizing new conducts (or new agents) we should ask ourselves: could administrative or civil law achieve the prevention of negative outcomes in an equally effective way? If yes, "which solution is less costly (for the state, for

[1263] Lagioia & Sartor, 2020, p. 7.
[1264] Lagioia & Sartor, 2020, p. 7.
[1265] Sarch, 2022, p. 311.

society and the economy, and offenders)?".[1266] As of now, the criminalization of AI systems appears to be the costliest solution. Things could change in the future, for example if we were to face AI systems which could pose grave threats to collective goods: this could entail that punishing AI system would be in compliance with the *ultima ratio* principle.

8.5.1 Looking Forward

In a perspective for the future, I agree with those who claim that the introduction of criminal liability of AI systems will depend on whether the society of the future will develop "normative expectations" towards them.[1267] If AI systems are expected to comply with criminal norms, then their noncompliance will require the reaction of criminal law.[1268] As mentioned above, this attribution could be mandated, for example, by a change in the values of our community, or by a development in the technology which will lead to fully "humanized" robots.

If we were to decide that AI systems can actually commit crimes, further research could be pursued, also in light of this study's result, regarding whether restorative justice could "offer an alternative way of dealing with the occurrence of AI crimes".[1269] The topic has not (yet) gained much attention. Some have acknowledged that restorative justice "might support victims of AI crimes better than the punitive legal system, as it allows for the sufferers of AI crimes to be heard in a personalised way"[1270] and that it would be more affordable for victims, since "AI technologies would not only be defined by a powerful elite".[1271] Surely, this opens up the question of how reconciliation between AI systems and victims would take place.

As it was claimed, intent and, more generally, criminal capacity, which are lacking at the moment in AI systems, denote being susceptible to the command enshrined in a criminal norm. Yet, some contend that the criminal norm includes more than deterrence achieved through the threat of the infliction of suffering.[1272] In view of that, it is argued that criminal

---

[1266] Hörnle, 2014, p. 696.
[1267] Ivi, p. 25.
[1268] Ibid.
[1269] A. Hadzi & D. Roio, "Restorative Justice in Artificial Intelligence Crime", *Spectres of AI,* Vol. 5, 2019, p.16.
[1270] Ivi, p. 17.
[1271] Hadzi & Rio, 2019, p. 17.
[1272] Colucci, 2022; Papa, 2019.

offenses have a "guidance" and "communicative" function, which is usually best achieved by communicating with its recipients via the prospect of a reward (such as that of not being punishable for a previous illegal action).[1273] This is relevant when dealing with AI systems which, thanks to reinforcement learning techniques, can be trained with a reward-penalty method.

Imagine that, in the future, we were to face the production of rules that can, from the very beginning, be consumed directly by machines.[1274] The very day that legal rules will lose certain characteristics (such as the reference to sensitive reality), and become very formal (to the point that they can be expressed by bar codes),[1275] criminal compliance by AI systems will stop being different from that of humans. This might change our perspective on criminal capacity altogether.

Furthermore, one option could be to introduce punishment of AI systems *only for specific* (existing) crimes. In fact, modern criminal law has shifted from punishing the so-called *malum in se* offenses (i.e., offenses that are inherently bad, evil, or morally wrong), to *malum prohibitum* offenses (offenses that are wrong only because they are prohibited by positive laws).[1276] One of the most prominent examples is the field of environmental criminal law, where harm results from "accumulative acts which, at the individual level, would be harmless".[1277] Moreover, it is assumed that contemporary criminal law should be focused on behavior and not on thoughts, following the *cogitationis poenam nemo patitur* maxim.[1278]

This could work similarly as in those legal systems where corporations are punished according to a "catalogue" of offenses which very often includes an extensive list of *malum prohibitum* offenses. Such offenses present a very "thin" *mens rea*, due to their intrinsic

---

[1273] Colucci, 2022, pp. 9 and 201; M. Papa, *Fantastic Voyage,* 2nd Ed., Giappichelli, 2019, pp. 108 ff.

[1274] Papa, 2019, pp. 243-244.

[1275] Ibid.

[1276] " 'Crimes mala in se' embrace acts immoral or wrong in themselves, such as burglary, larceny, arson, rape, murder, and breaches of peace. […] 'Crimes mala prohibita' embrace things prohibited by statute as infringing on others' rights, though no moral turpitude may attach, and constituting crimes only because they are so prohibited". See H. C. Black, *Black's Law Dictionary,* 4th Ed, West Publishing, 1968, pp. 445-446. For a unique reflection on the crisis of offense definitions in contemporary criminal law see Papa, 2019; M. Papa, "The Offense Definition as a Screenplay of Evil: The Rise and Fall of Visual Criminal Law", *Católica Law Review*, Vol. 4, No. 3, 2020, pp. 145-174.

[1277] N. Peršak, *The Harm Principle, its Limits and Continental Counterparts*, Springer, 2007, p.137. Other less eminent examples can be found in the United States Code. See J. G. Malcolm, "Morally Innocent, Legally Guilty: The Case for Mens Rea Reform", *The Federalist Society Review*, Vol. 18, 2017, p.42.

[1278] Digest of Justinian, 48.19.18, Ulp. 3 ad ed.

regulatory nature, hence would pose less problems with regards to the attribution of a sufficient level of "disregard", i.e., intent, to AI systems. Usually, one of the concerns regarding statutory offenses is that they are highly technical and "hidden" in *extra-codicem* regulations, so that they can lead to punishment of individuals who were unaware of their existence (according to the general *ignorantia legis non excusat* principle). This issue might not be as pressing for AI systems, since they are capable of reading and memorizing incredible amounts of texts by using natural language processing.[1279] Nevertheless, if awareness of the norm entails not only the knowledge of its existence, but also the comprehension of its demands, then we are back to square one. Moreover, one could think of introducing new AI-specific offenses, such as a commission-by-omission offense connected to a duty to save human lives (think of the example of the heroic robot discussed throughout this research).[1280] Definitely, such an expansion of the scope of application of criminal law should not happen as if criminal law were the *prima ratio* solution vis-à-vis a population's sentiment.[1281]

Surely, future investigations will have to take into account that folks attribution of criminal liability to AI systems seems to be going in the opposite direction of this research's results. As a matter of fact, recent research has shown that people apply different moral norms to human and robot agents, specifically, robots are more strongly expected to take a utilitarian choice, i.e., one that sacrifices one person for the good of many, in a moral dilemma

---

[1279] Research on the matter, also referred to as "LegalAI", is vast. See *ex multis* H. Zhong & al., "How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5218–5230; M. J. Bommarito II, D. M. Katz & E. M. Detterman, "LexNLP: Natural language processing and information extraction for legal and regulatory texts", in R. Vogl (Ed)., *Research Handbook on Big Data Law,* Edward Elgar Publishing, 2021, pp. 216-227; I. Chalkidis, M. Fergadiotis & I. Androutsopoulos, "MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer", in M. Moens et al. (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 2021; I. Chalkidis & al., "LEGAL-BERT: The Muppets straight out of Law School", *Findings of EMNLP*, 2020; D. Tsarapatsanis & N. Aletras, "On the Ethical Limits of Natural Language Processing on Legal Text", *ACL-IJCNLP*, 2021, pp. 3590–3599.

[1280] See above at 1.6 and at 8.4.1.

[1281] See Peršak, who underlines the importance of the principle of harm with regards to criminalization: "The principle, namely, sets out the legitimate grounds, necessary conditions that need to be fulfilled at the first stage of criminalisation, and the condition is 'harm to others' – not 'immorality', not harm to self (and in the extreme liberal version also no offence), only to others. The harm principle, […] if supported by rules of fair imputation and various limiting principles, should also militate against (or, *ex post*, declare illegitimate) the criminalisation on the basis of mere abstract, far-fetching risk or remote harm, which often covers others (illegitimate) reasons disguised in the rhetoric of harm." Peršak, 2007, p.135.

situation, and are blamed more than their human counterparts when they do not do so.[1282] Hence, in the example of the robot and the drowning child, the robot is expected, and might even be obliged, to save the child no matter what.[1283]

Besides, humans seems to be inclined towards blaming AI systems for omissions more than humans, where instead the trend goes in the opposite direction when it comes to actions (i.e., humans are blamed more than AI systems for a wrongful act).[1284] What is more, it has also been shown that people consider AI systems as capable of displaying *mens rea*[1285] and, in fact, are inclined to attribute *mens rea* states (specifically, "artificial recklessness")[1286] to AI systems, sometimes as much as they do to human agents or to corporations.[1287] It must be discussed whether this type of research will have an impact on criminal policy decisions regarding whether criminal liability of AI systems is *needed*. Those who argue that the bedrock of *mens rea*, i.e., free will, is nothing but a matter of social attribution, might answer yes.[1288] This (open) question is perhaps one of the most stimulating outcomes of this research.

8.5.1 Looking Backwards

Let us conclude now with a few final reflections on the trajectory that led to this studies results.

It has been argued that

> Clever scholars can adjust existing criminal theories to the AI field and make them fairly plausible. However, what we are agonizing over is not whether we can conjure

---

[1282] B. F. Malle et al., "Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents", *HRI '15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction,* 2015, pp. 117-124.

[1283] Ibid.

[1284] Malle et al., 2015; B. F. Malle, S. T. Magar & M. Scheutz, "AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma", in M. I. Aldinhas et al. (Eds.), *Robotics and Well-Being. Intelligent Systems*, Control and Automation: Science and Engineering, Vol. 95, Springer, 2019, pp. 111-133; Malle et al, "Which Robot Am I Thinking About? The Impact of Action and Appearance on People's Evaluations of a Moral Robot", *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, 2016, pp. 125–132.

[1285] M. Kneer & M. T. Stuart, "Playing the Blame Game with Robots", *HRI '21 Companion: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 408 ff.

[1286] Kneer & T. Stuart, 2021, p. 408

[1287] Ibid.

[1288] Simmler & Markwalder, 2019, p. 15.

> up a decent plausible theory to justify criminal liability, but whether AI is really liable and its punishment itself is justifiable.[1289]

This quote pinpoints one of the most pressing difficulties of this research: the vastity of the topic addressed. Certainly, I did not intend to reinvent the wheel (i.e., at writing a new general theory of criminal law), even though sometimes it felt as burdensome. Indeed, an analysis of the impact of AI on substantial criminal law has potentially endless ramifications. For instance, think of AI and criminal capacity (Chapter 5): exploring whether AI could be a subject of criminal law presupposed addressing the lack of understanding of what an AI "agent" is, but also acknowledging that the concept of criminal capacity per se is not recognized in every legal system; moreover, it also involved dealing with the fact that, provided that one accepts that criminal capacity as a legal notion even exists, there is no agreement on what it is made of.

The ramifications of this research did not limit themselves to different theories and constructs of criminal law, but also to other disciplines: if there is one lesson which could be drawn from this research is that it requires criminal law scholars to step out of their comfort zones. If one wants to deal with criminal liability of AI system, and wants to do so effectively, she is compelled to acquire notions of computer science, to truly grasp the functioning of AI systems; she is drawn to theories of moral philosophy, to understand purposes of punishment in connection to AI; and she has to become acquainted with ethics, due to the importance and prominence of ethical guidelines and principles as forebearers of AI regulation; and so forth. This complexity caused the need to constantly draw boundaries delineating what was part of the research, and what was not.

Furthermore, the strength (and the weakness) of this research is that it is deeply *introspective*: it prompts questions which find their answer in one's own beliefs on what is human and what is not, and, ultimately, on what is right and what is wrong. As such, most of the questions asked (and answered) in this research could be seen as controversial by the readers. When addressing these consequential questions, the research involved a great deal of "reasoning by proxy", that is, it compared AI systems to other subjects that displayed similar characteristics, be it animals, insane offenders, infants, or corporations. Thus,

---

[1289] Lee, 2021, p. 321.

reasoning by proxy must be handled with care: if one is not careful, it could turn into trying to square a peg in a round hole.

In conclusion, as this research has shown, considering the criminal behavior and accountability of AI systems is not (just) a matter of legal fictions and of academic discussions. In some cases, the decisions and the actions of such systems may cause harm, or the serious threat of it, leading to the question of who should be held responsible.

Overall, I believe that discussing criminal liability of AI systems is essential for addressing the legal and ethical challenges posed by the use of AI and ensuring that AI systems are developed and used in a responsible and accountable manner.

Or did an AI system just write that?[1290]

[1290] The answer is yes. The last paragraph was written by ChatGPT, in answer to my question "What do you think of the importance of discussing the issue of criminal liability of AI systems?". ChatGPT is a conversational chatbot (based on natural language processing) developed by OpenAI. It can be accessed at: chat.openai.com.

SUMMARY

The research deals with the subject of criminal responsibility *of* Artificial Intelligence (AI) systems, focusing on whether such a legal framework is *needed* and *feasible*.

Chapter 1 presents the main RQ (to what extent is a theoretical framework of criminal liability for non-human agents *needed* and *feasible*?) and the issues that that will be discussed throughout the research, together with a structure of the corresponding sub-questions. It then outlines the methodology of the study and provides a set of examples of "AIs going bad".

Chapter 2 tackles the issue of defining AI and adopts the AI-HLEG definition as working definition for the study. This provides foundation to the analysis. The definition will then be tested throughout the following chapters and assessed in Chapter 8.

Chapter 3 delivers an extensive literature review, which is used to situate the study amongst other scholarly outputs. The analysis of scholarly debate on AI and criminal law is based on over 100 sources written in three languages (Italian, English, and German). The authors are divided into three categories: expansionists, moderates, and skeptics. The Chapter is concluded with the identification of 10 recurring questions and 7 gaps.

Chapter 4 introduces the heart of the study: an analysis which mirrors, in structure, the classical construct of criminal offenses. Indeed, AI seems to clash with traditional notions of criminal law, and understanding how to do deal with this (apparent) conflict is one of the research's tenets. In order to discuss said issues, the Chapter presents an analogy between the field of AI and that of aviation, together with an overview of different theories of criminalization.

Chapter 5 focuses on whether AI systems could display the prerequisites of criminal liability, i.e., the characteristics that are needed in order for a subject to be a plausible addressee of a criminal norm. Such a reflection is conducted by discussing the connection between moral

and illegal wrongs and by examining whether AI systems could be considered moral and/or legal agents. This Chapter advances the idea that the *comprehension* of an offense's command is an essential prerequisite for establishing criminal liability and that, as such, it represents the main obstacle to attributing liability to AI systems directly.

Chapter 6 considers whether AI behavior could fulfil the *mens rea* requirement of a criminal offense, i.e., whether an AI system could be deemed "guilty". When doing so, it also looks at humans-behind-the-machine. Specifically, it addresses situations in which the classical building blocks of negligence, i.e., risk taking, foreseeability, and awareness, struggle to identify a liable human being to whom we can attribute AI-caused harm. Then, the chapter shifts its focus to whether AI behavior could fulfil the *actus reus* requirement. In particular, it identifies three main issues which obstruct the identification of a clear causal nexus between an AI act and the realization of harm. Subsequently, it analyzes with a critical eye the differences and similarities between AI criminal liability and corporate criminal liability.

Chapter 7 provides an outline of the state of the art regarding the adoption of hard and/or soft law instruments directed at regulating AI and criminal liability. In particular, it analyzes: A – The Council of Europe's European Committee of Criminal Problems and the drafting of an "Instrument on Artificial Intelligence and Criminal Law"; B – the Singapore Penal Code Review Committee Report of 2018 and the Report on "Criminal Liability, Robotics and AI systems" drafted by the Singapore Law Commission of 2021; C – the legislative reform of the French Road Act (*Ordonnance n° 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation*); D – the "Automated Vehicles: joint report" drafted by the Law Commission of England and Wales and by the Scottish Law Commission; E – the amendment of the German Road Traffic Act.

Chapter 8 retraces the questions, and the interim conclusions, posed throughout the thesis, and identifies avenues for further inquiry. It suggests that the two prongs of the research question are switched, i.e., that the RQ is reformulated as follows: to what extent is a theoretical framework of criminal liability for non-human agents feasible and needed. The succinct answer to the first prong of the RQ is the following: criminal liability of AI agents

is *feasible* depending on how much of our anthropomorphic notions we are willing to give up. Since criminal liability can be imputed only to individuals who are capable of understanding and following the *demands* of criminal law – i.e., to make a deliberate choice *not* to conform to normative expectations, and since it is not possible, as of now, to attribute such capacity to AI systems, holding AI systems responsible would entail either being able to create machines that are *susceptible* to the commands of a legal norm, or abandoning the concept of rationality (meant as the ability to be responsive to the law) altogether. The succinct answer to the second prong of the RQ is the following: criminal liability of AI agents is not *needed* as long as AI systems are not treated as members of our community and, as such, as subject to the same rules as us. The change in perception, i.e., the anthropomorphization of AI Systems, could happen only the moment "true" Artificial Moral Agents (AMAs) are developed. Being an AMA would suffice to pin moral blame on AI systems. Yet, in order to be considered criminally culpable, AI systems should also be "legal" agents. And, in other to be legal agents, AI systems would have to be responsive to criminal norms. In other words, the questions of whether criminal liability of AI systems is feasible, and needed, remained intertwined.

SINTESI

La ricerca affronta il tema della responsabilità penale *dei* sistemi di Intelligenza Artificiale (IA), concentrandosi sulla sua *necessità* e *fattibilità*.

Capitolo 1. Il primo capitolo introduce la domanda di ricerca principale (*To what extent is a theoretical framework of criminal liability for non-human agents* needed *and* feasible?) e le tematiche che saranno discusse nel corso della ricerca, insieme a una sintesi delle sotto-domande correlate ai vari capitoli. La ricerca illustra poi la metodologia dello studio e fornisce una serie di esempi di "*AIs going bad*".

Capitolo 2. Per dare fondamento all'analisi, lo studio inizia affrontando la questione della mancanza di una definizione universalmente accettata di IA e adotta la definizione fornita dall'AI-HLEG quale definizione operativa per la ricerca. Tale definizione viene poi testata nei capitoli successivi e la sua adeguatezza viene valutata nel Capitolo 8.

Capitolo 3. Il capitolo offre un'ampia rassegna della dottrina in materia di IA e diritto penale, utilizzata quale base per collocare questa ricerca all'interno del dibattito. L'analisi si basa su oltre cento fonti scritte in tre lingue (italiano, inglese e tedesco). Gli autori sono suddivisi in tre categorie: espansionisti, moderati e scettici. Il capitolo si conclude con l'individuazione di dieci domande ricorrenti e sette lacune.

Capitolo 4. Il capitolo introduce il cuore dello studio: l'impatto dell'IA sui costrutti classici del diritto penale. Per fare ciò, viene proposta un'analogia fra il campo dell'IA e quello dell'aviazione, nonché una disamina delle analisi delle principali teorie poste alla base delle scelte di criminalizzazione.

Capitolo 5. Il capitolo si concentra sull'imputabilità dei sistemi di IA, ossia se questi possano possedere le caratteristiche necessarie affinché vengano trattati quali "destinatari plausibili" della sanzione penale. Tale riflessione viene condotta approfondendo la connessione tra illecito morale e illecito penale e, di conseguenza, viene discusso se i sistemi di IA possano

essere considerati agenti "morali" e/o soggetti attivi ai fini della commissione di un reato (c.d. "*legal agents*"). In questo capitolo viene presentata la tesi secondo cui l'incapacità dei sistemi di IA di *comprendere* il comando espresso dalla fattispecie penale rappresenterebbe il principale ostacolo all'attribuzione diretta di responsabilità penale in capo ad essi.

Capitolo 6. Il capitolo esamina se il comportamento dell'IA possa soddisfare il requisito dell'elemento soggettivo di un reato, ossia se un sistema IA possa essere considerato "colpevole". Nel fare ciò, viene posta attenzione anche ai c.d. "*humans-behind-the-machine*". In particolare, per quanto riguarda quest'ultimi, viene discussa la possibilità di configurare una responsabilità a titolo colposo a loro carico. Uno degli snodi più problematici si rinviene nel livello di attenzione esigibile dal potenziale supervisore a causa di alcuni fenomeni diffusi nel campo dell'automazione, quali ad esempio l'*automation complacency* e l'*automation bias*. Tali fenomeni, amplificati nel caso di automatizzazione basata su IA, riducono sensibilmente la soglia di attenzione dell'agente umano alle prese con la macchina, e, di conseguenza, diminuiscono la soglia dell'esigibilità di una condotta diligente da parte dello stesso.
Il capitolo analizza poi se l'agire dell'IA possa essere considerato rilevante per la qualificazione dell'elemento oggettivo del reato. Per quanto riguarda l'accertamento del nesso causale, in particolare, vengono individuati tre ostacoli principali: il "*problem of many hands*", il problema della "*black box*" e il problema delle "scorciatoie". Successivamente il capitolo analizza con occhio critico le differenze e le affinità tra la responsabilità penale dell'IA e la responsabilità penale delle persone giuridiche.

Capitolo 7. Questo capitolo fornisce una panoramica dello stato dell'arte relativo all'adozione di strumenti di *hard* e/o *soft law* volti a regolamentare l'IA e la responsabilità penale. In particolare, analizza: A – la stesura di uno "Strumento sull'intelligenza artificiale e il diritto penale" da parte del Consiglio d'Europa; B – il rapporto del Comitato di Revisione del Codice Penale di Singapore del 2018 e il rapporto su "Responsabilità penale, robotica e sistemi di intelligenza artificiale" redatto dalla Singapore Law Commission nel 2021; C – la riforma legislativa del codice della strada francese; D – il "*Automated Vehicles: joint report*" redatto dalla Law Commission di Inghilterra e Galles e dalla Law Commission scozzese; E – la modifica del codice stradale tedesco.

Capitolo 8. Questo capitolo ripercorre le domande e le conclusioni intermedie poste nel corso della tesi e identifica percorsi per future indagini. Suggerisce di scambiare i due elementi su cui si articola la domanda di ricerca principale, ossia di analizzare prima la fattibilità di un meccanismo di imputazione di responsabilità penale diretta in capo ai sistemi di IA, e interrogarsi dopo sulla sua necessità. La risposta sintetica alla prima parte della domanda di ricerca è la seguente: la responsabilità penale degli agenti algoritmici è possibile a seconda di quanto siamo disposti a rinunciare alle nostre nozioni giuridiche antropomorfiche. La responsabilità penale può essere attribuita solo a coloro che sono in grado di comprendere, e aderire, alle "richieste" del diritto penale – soggetti cioè capaci di fare una scelta intenzionale di non conformarsi alle aspettative normative. Ad oggi, non è possibile attribuire tale capacità ai sistemi di IA. Ne consegue che al fine di poter ritenere i sistemi di IA penalmente responsabili sarebbe necessario o essere in grado di creare macchine che siano suscettibili ai comandi di una norma giuridica, o abbandonare del tutto il concetto di razionalità (intesa come capacità di "rispondere alla legge"). La risposta sintetica alla seconda parte della domanda di ricerca è la seguente: la responsabilità penale dei sistemi di IA non è necessaria finché i sistemi di IA non verranno ritenuti membri della nostra comunità e, in quanto tali, soggetti alle nostre stesse regole. Il cambiamento di percezione, cioè l'antropomorfizzazione dei sistemi di IA, potrà avvenire ove vengano sviluppati dei "veri" agenti morali artificiali. Tuttavia, sebben essere un "agente morale artificiale" sarebbe sufficiente per attribuire una colpa morale ai sistemi di IA, non sarebbe adeguato a considerare gli stessi agenti penalmente colpevoli. Difatti, per essere considerati soggetti attivi di un reato, i sistemi di IA dovrebbero rispondere al comando delle fattispecie di diritto penale e ciò, ad oggi, non è ancora possibile. In altre parole, fattibilità e necessità di un quadro di responsabilità penale dei sistemi di IA rimangono intrecciate.

SAMENVATTING

Het in dit proefschrift gepresenteerde onderzoek gaat over de strafrechtelijke aansprakelijkheid van kunstmatige intelligentie-systemen (AIs).

Hoofdstuk 1 presenteert de hoofdvraag: in hoeverre is een theoretisch kader van strafrechtelijke aansprakelijkheid van niet-menselijke actoren noodzakelijk en mogelijk? Daarnaast worden de in dit proefschrift neergelegde fundamentele kwesties geïntroduceerd, inclusief de corresponderende deelvragen. Het introductiehoofdstuk eindigt met een presentatie van de methodologie waar het proefschriftonderzoek op is gebaseerd en enkele illustratieve voorbeelden van "AIs going bad".

Hoofdstuk 2 Gaat in op de kwestie hoe AI te definiëren. Concreet wordt er daarbij aangesloten bij de AI-HLEG-conceptualisering. In de resterende hoofdstukken van het proefschrift wordt deze definitie getoetst en in hoofdstuk 8 op zijn merites beoordeeld.

Hoofdstuk 3 situeert het in dit proefschrift gepresenteerde onderzoek binnen de volle reikwijdte van het wetenschappelijke debat over AI en strafrecht. Daarvoor zijn er meer dan 100 publicaties geraadpleegd in drie talen (Italiaans, Engels en Duits). De respectievelijke auteurs en hun wetenschappelijke productie wordt gegroepeerd in drie categorieën: expansionisten, gematigden en sceptici. Het hoofdstuk wordt afgesloten met de signalering van tien terugkerende vragen en zeven lacunes.

Hoofdstuk 4 presenteert het fundament van het onderzoek: een analyse die qua structuur de klassieke indeling van strafbare feiten volgt. AI lijkt immers niet in traditionele strafrechtterminologie te positioneren: een van de uitgangspunten van dit onderzoek is hoe met dit (schijnbare) conflict moet worden omgegaan. In dit hoofdstuk wordt daarom AI en de luchtvaart vergeleken en een overzicht gepresenteerd van verschillende theorieën over criminalisering.

Hoofdstuk 5 gaat in op de vraag of AI-systemen kunnen voldoen aan de voorwaarden voor strafrechtelijke aansprakelijkheid. Concreet: of AI-systemen de kenmerken herbergen voor het 'zijn' van een adressant van een strafrechtelijke norm. Deze vraag wordt beantwoord aan de hand van een presentatie van het verband tussen morele en illegale misstanden en door na te gaan of AI-systemen kunnen worden beschouwd als morele en/of rechtspersonen. In dit hoofdstuk wordt het idee verdedigd dat begrip van een strafbaar feit een essentiële voorwaarde is voor de vaststelling van strafrechtelijke aansprakelijkheid en dat dit als zodanig het belangrijkste obstakel vormt om AI-systemen rechtstreeks aansprakelijk te stellen.

Hoofdstuk 6 beschouwt of AI-gedragingen kunnen voldoen aan de mens rea-eis van een strafbaar feit. In het kort: kunnen AI-systemen "schuldig" zijn? Daarbij wordt gekeken naar de mens achter de betreffende AI-machines. In het bijzonder wordt ingegaan op situaties waarin de klassieke bouwstenen van nalatigheid, te weten risicobereidheid, voorzienbaarheid en besef, het moeilijk maken om een aansprakelijk mens aan te wijzen aan wie door AI veroorzaakte schade kan worden toegerekend. Vervolgens gaat het hoofdstuk in op de vraag of AI-gedragingen kunnen voldoen aan het actus reus-vereiste. In het bijzonder worden de drie belangrijke kwesties geïdentificeerd die de identificatie van een evident causaal verband tussen een AI-handeling en het ontstaan van schade blokkeren. Vervolgens worden kritisch de verschillen en overeenkomsten tussen strafrechtelijke aansprakelijkheid voor AI en strafrechtelijke aansprakelijkheid van ondernemingen gepresenteerd.

Hoofdstuk 7 geeft een overzicht van de stand van zaken met betrekking tot de opname van hard en/of soft law-instrumenten ter regulering de strafrechtelijke aansprakelijkheid van AI. In het bijzonder wordt ingegaan op A – het Europees Comité voor strafrechtelijke vraagstukken van de Raad van Europa en het "Instrument on Artificial Intelligence and Criminal Law"; B – the Singapore Penal Code Review Committee Report of 2018 en het rapport over "Criminal Liability, Robotics and AI systems", zoals opgesteld door de Singapore Law Commission of 2021; C – de herziening van de Franse Wegenwet (*Ordonnance n° 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation*), D – de "Automated Vehicles: joint report" opgesteld door de Law Commission of England and Wales en de Scottish Law Commission; E – het amendement van de Duitse *Straßenverkehrsgesetz*.

Hoofdstuk 8 bespreekt de vragen en deelconclusies die in het proefschrift zijn opgeworpen nog een laatste maal. Daarnaast worden enkele aanbevelingen voor verder onderzoek gepresenteerd. Zo wordt voorgesteld om de twee onderdelen van de onderzoeksvraag om te wisselen. De vraag leest dan als volgt: in hoeverre is een theoretisch kader van strafrechtelijke aansprakelijkheid voor niet-menselijke agenten haalbaar en nodig. Het beknopte antwoord op het eerste deel van de vraag: strafrechtelijke aansprakelijkheid van AI-agenten is mogelijkheid, maar afhankelijk van de mate waarin wij bereid zijn onze traditionele antropomorfe opvattingen los te laten. Aangezien strafrechtelijke aansprakelijkheid alleen kan worden toegeschreven aan individuen die in staat zijn de eisen van het strafrecht te begrijpen en te volgen – i.e. een afgewogen keuze te maken om zich niet aan zekere normatieve verwachtingen te conformeren - en aangezien het op dit moment niet mogelijk is een dergelijk vermogen aan AI-systemen toe te kennen, zou het verantwoordelijk stellen van AI-systemen betekenen dat men ofwel in staat moet zijn machines te creëren die ontvankelijk zijn voor de bevelen van een wettelijke norm, ofwel het begrip rationaliteit (in de zin van het vermogen om te reageren op het recht) helemaal moet worden opgeven. Het beknopte antwoord op het tweede punt van de hoofdvraag luidt als volgt: strafrechtelijke aansprakelijkheid van AI-agenten is niet nodig zolang AI-systemen niet worden behandeld als leden van onze gemeenschap die aan dezelfde regels zijn onderworpen als wij. Een verandering in perceptie, i.e. de antropomorfisering van AI-systemen, zou pas plaats kunnen vinden op het moment dat er "echte" Artificial Moral Agents (AMA's) worden ontwikkeld. Het 'zijn' van een AMA zou het dan mogelijk maken om AI-systemen moreel te beschuldigen. Maar om strafrechtelijk verwijtbaar te zijn, moeten AI-systemen ook rechtspersonen zijn. En om rechtspersonen te zijn, zouden AI-systemen moeten reageren op criminele normen. Met andere woorden, de vraag of strafrechtelijke aansprakelijkheid van AI-systemen haalbaar en nodig is, blijft met elkaar verweven.

IMPACT STATEMENT

SOCIETAL AND SCIENTIFIC RELEVANCE OF THE RESEARCH FINDINGS

Questions of accountability and liability for the actions of AI systems will surface, as they become more sophisticated and common. As such, this study could lead to the development of legal frameworks that bridge accountability gaps, bearing significant implications for industries such as healthcare, finance, and transportation. Moreover, this study could lead to a change in public perception of crimes committed by AI.

Indeed, one of the main results obtained with this research is that it is able to explain the impact of AI on matters of criminal law to non-lawyers. The language adopted is purposedly straight-forward and simple, hence the research can be read also by those who do not have a legal education. In this sense, it could work to build bridges in the future between criminal legal scholars and AI scientists. Part of this result were already obtained, via presenting the research findings to a network comprising of both AI scholars and of philosophers. Furthermore, some of the reflections were already anticipated in publications on legal journals (especially on the matter of liability of humans).

In the short-term, the results of this research definitely brought the discussion on the matter forward: as a *unicum* in its field, it is hoped that it will stimulate further research on the matter. Specifically, this research will be useful for legislators and policymakers when drafting new legal tools on matters of AI and criminal law. It could also impact international policy makers, who may look at the research to inform their own policies on AI systems. In this sense, its wide scope and lack of boundaries to national legal systems proves extremely useful.

INNOVATIVE ASPECTS OF THE RESEARCH

The study provides an original point of view in a doctrinal debate which has become topical in the past 5 years. The originality stems from different aspects. First and foremost, it represents a wholesome synthesis of a complex and layered topic. Indeed, an analysis of the impact of AI on substantial criminal law has potentially endless ramifications: in order to be able to truly deal in depth with the issue of criminal liability of AI system, a researcher has to acquire notions of computer science (so as deal with the functioning of AI systems in laymen terms); notions related to theories of moral philosophy (so as to deal with purposes of

punishment in connection to AI); and notions related to ethics (due to the importance and prominence of ethical guidelines and principles as forebearers of AI regulation). The book is the result of the combination of these notions in a way that is not redundant and understandable for legal scholars. Second, it delivers – for the first time – an extensive analysis of the scholarly debate on AI and criminal law based on over 100 sources written in 3 languages (Italian, English and German). Third, it presents reflections on avant-garde topics, which have not been scrutinized by scholars before (see for example the discussion on Artificial Insane Offenders and Artificial Infant Offenders and the analysis of duties to act and responsibility of AI systems in commission-by-omission scenarios). Fourth, it introduces a unique comparison of 5 novel legal frameworks (laws, law proposals, and other policing initiatives) related to criminal liability and AI systems. Finally, the conclusion chapter presents numerous open questions, which will potentially be used by researchers as inputs for future research.

TARGET AUDIENCE & OUTREACH

The main target group for this research are criminal legal scholars interested in the interaction between criminal law and technology. Thus this study, due to its polyhedric nature, is of interest also for AI scientists, ethicists, and philosophers. It provides a basis for further discussion with academics from various fields. Thus, the research is not only relevant for members of academia, but also for member of parliaments, governmental officials, and of other public bodies, such as law enforcement agencies. For example, legislators could make use of this research when faced with the issue of regulating criminal behavior of AI systems. In fact, part of the research' results were already presented to high-ranking members of law enforcement agencies enrolled in a master on AI in criminal justice. Furthermore, this research could also affect the technological advancement of AI systems, stimulating industries to include considerations on criminal liability in their production processes. Finally, in the future the findings of the study could be integrated into education, e.g., by including tailored lectures in courses at universities. The lectures could be given to students belonging to different faculties (e.g., law, philosophy, data science, mathematics) and foster interdisciplinary research in the future.

INDEX OF FIGURES

Figure 1	An example of POET'S bipedal-walking obstacle course.	Paragraph 2.1, p. 27
Figure 2	Artificial intelligence over the years	Paragraph 2.2, p. 31
Figure 3	Expert Systems	Paragraph 2.2., p. 33
Figure 4	Functioning of an Artificial Neural Network	Paragraph 2.3, p. 36
Figure 5	Relationship between AI, ML, ANNs and DL	Paragraph 2.3, p. 37
Figure 6	Segmentation of general object recognition	Paragraph 2.3, p. 38
Figure 7	Example of a question asked during the MIT'S Moral Machine Experiment	Paragraph 5.3.2.1, p. 146
Figure 8	Comparing the results of the MIT experiment of Germans (green) and Americans to the world average (grey).	Paragraph 5.3.2.1, p. 147
Figure 9	The Burning Room Dilemma	Paragraph 5.3.2.2, p. 148
Figure 10	Definitions of intent	Paragraph 6.2.2, p. 176

BIBLIOGRAPHY

DOMESTIC LEGISLATION

FRANCE

Code de la route.

Décret n° 2021-873 du 29 juin 2021 portant application de l'ordonnance n° 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation, NOR : TRAT2034544D.

Ordonnance n° 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation (TRAT2034523R, JORF n°0089 du 15 avril 2021, Texte n° 36), 2021.

GERMANY

Bundestag, Gesetz zur Änderung des Straßenverkehrsgesetzes und des Pflichtversicherungsgesetzes – Gesetz zum autonomen Fahren, Bundesgesetzblatt Jahrgang 2021, Teil I Nr. 48, 3108.

UNITED KINGDOM

Automated and Electric Vehicles Act, 2018.

USA

Arizona Revised Statutes, 1956.

Code of Federal Regulations (CFR), National Archives and Records Administration's (NARA) Office of the Federal Register (OFR), and the Government Publishing Office (GPO).

DOMESTIC POLICY PAPERS, RESEARCH DOCUMENTS, AND REPORTS

GERMANY

Ethics Commission Automated and Connected Driving, 2017, "Report". Available at: https://bmdv.bund.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile.

SINGAPORE

Singapore Academy of Law, Law Reform Committee, *Report on Criminal Liability, Robotics and AI Systems*, 2021. Available at: https://www.sal.org.sg/sites/default/files/SAL-LawReform-Pdf/2021-02/2021%20Report%20on%20Criminal%20Liability%20Robotics%20%26%20AI%20Systems.pdf.

Singapore Penal Code Review Committee, *Report*, 2018. Available at: https://www.reach.gov.sg/-/media/reach/old-reach/2018/public-consult/mha/annex--pcrc-report.ashx.

UNITED KINGDOM

"Automated Vehicles: Consultation Paper 3 – A regulatory framework for automated vehicles A joint consultation paper", Law Commission Consultation Paper No 252, Scottish Law Commission Discussion Paper No 17. Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2021/01/AV-CP3.pdf.

Law Commission of England and Wales Law Commission No. 404, Scottish Law Commission Scottish Law Commission No. 258, "Automated Vehicles: Joint report", HC 1068 SG/2022/15, 25 January 2022 ("Law Commissions Report"). Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2022/01/Automated-vehicles-joint-report-cvr-03-02-22.pdf.

Law Commission of England and Wales Law Commission No. 404, Scottish Law Commission Scottish Law Commission No. 258, "Automated Vehicles: Summary of joint report", Summary of LC Report No 404 / SLC Report No 258, HC 1068 SG/2022/15, 26 January 2022. Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2022/01/AV-Summary-25-01-22-2.pdf.

USA

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., 24 May 1962.

National Safety Transportation Board (NTSB),

Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida, May 7, 2016, Highway Accident Report NTSB/HAR-17/02, 2017. Available at: https://www.ntsb.gov/investigations/accidentreports/reports/har1702.pdf.

Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian,Tempe, Arizona, March 18, 2018, Highway Accident Report NTSB/HAR-19/03, 2019. Available at: https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf.

National Highway Traffic Safety Administration, Office of Defects Investigation, Investigation PE 16-007, 2017.Available at: https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF.

United States Attorneys' Manual § 9-16.325, *Plea Agreements, Deferred Prosecution Agreements, Non-Prosecution Agreements and Extraordinary Restitution*, 2008.

INTERNATIONAL AND EUROPEAN LEGISLATION

European Commission, *Proposal for a Directive of the European Parliament and of the Council On Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence* (AI Liability Directive), COM(2022), 496 final 2022/0303 (COD), 28 September 2022.

European Commission, *Proposal for a regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence ("Artificial Intelligence Act") and Amending Certain Union Legislative Acts*, COM(2021) 206 final, 21 April 2022.

European Parliament, *Directive 2014/57/EU of the European Parliament and of the Council of 16 April 2014 on criminal sanctions for market abuse*.

European Parliament, Resolution "*Framework on the Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies*", 2020/2012(INL), 20 October 2020.
European Parliament, *Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103(INL)), 16 February 2017.
European Parliament, *Resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence*, (2020/2014(INL)), 20 October 2020.
International Civil Aviation Organization (ICAO), *Aircraft Accident and Incident Investigation*, Annex 13 to the Convention on International Civil Aviation.

INTERNATIONAL AND EUROPEAN POLICY PAPERS, RESEARCH DOCUMENTS, AND REPORTS

Boucher P. N., European Parliamentary Research Service - Scientific Foresight Unit, "What if AI regulation promoted innovation?"; PE 729.515, 2022, p. 2.
Council of Europe, CAHAI Secretariat, "Towards a regulation of AI systems", DGI (2020) 16. Available at: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.
Council of Europe - CEPEJ, *European Ethical Charter on the Use of AI in Judicial Systems Council of Europe*, 2018. Available at: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699.
Council of Europe, European Committee on Crime Problems (CDPC),
Concept Paper "Artificial intelligence and criminal law responsibility in Council of Europe member states - the case of automated vehicles", cdpc/docs 2018/cdpc(2018)14, 16 October 2018. Available at: https://rm.coe.int/cdpc-2018-14rev-artificial-intelligence-and-criminal-law-project-2018-/16808e64ad.
"Thematic session on artificial intelligence and criminal law. The approach in Council of Europe member states the case of automated vehicles", Programme, CDPC(2018)18, 28 November 2019. Available at: https://rm.coe.int/cdpc-2018-18-draft-programme-thematic-session-artificial-intelligence-/16808e64ab.
"Thematic session on Artificial Intelligence and Criminal Law of 28 November 2018", Final remarks by Professor Sabine Gless, Special Rapporteur, "Artificial intelligence and its impact on CDPC work. The case of automated driving", cdpc/docs 2018/cdpc (2018)22, 28 November 2018.
Working Group of Experts on Artificial Intelligence and Criminal Law, "Working paper II", CDPC(2019)7, 27 March 2019, p. 3. Available at: https://rm.coe.int/cdpc-2019-7-working-paper-ii-for-cdpc-expert-group-meeting-on-artifici/16809372a5.
Working Group of Experts on Artificial Intelligence and Criminal Law, "Questionnaire concerning Artificial Intelligence and Criminal Justice (using the example of Automated Driving)"' CDPC(2019)8FIN (2019), 19 May 2019. Available at: https://rm.coe.int/cdpc-2019-8fin-questionnaire-artificial-intelligence-and-criminal-just/168094c8fa
"Assessment of the answers to the questionnaire on artificial intelligence and criminal justice (using the example of Automated Driving)", CDPC(2019)17 (2019), 7 November 2019. Available at: <https://rm.coe.int/cdpc-2019-17-draft-assessment-of-the-answers-to-the-questionnaire-on-a/168098e24c>.
Working Group on AI and Criminal Law & CDPC Secretariat, "Feasibility study on a future Council of Europe instrument on artificial intelligence and criminal law",

CDPC(2020)3Rev (2020), 4 September 2020. Available at: https://rm.coe.int/cdpc-2020-3-feasibility-study-of-a-future-instrument-on-ai-and-crimina/16809f9b60.
"1st meeting of the Drafting Committee to elaborate an instrument on Artificial Intelligence and Criminal Law (CDPC-AICL)", CDPC-AICL(2021), 17 November 2021. Available at: https://rm.coe.int/cdpc-aicl-2021-1-1st-meeting-report-15-16-nov-2021/1680a49c99.
"Terms of Reference", Extract from CM(2021)131-addrev, p. 2. Available at: https://rm.coe.int/cdpc-en-terms-of-reference-cm-2021-131-addrev/1680a4b41a
"2nd meeting of the Drafting Committee to elaborate an instrument on Artificial Intelligence and Criminal Law (CDPC-AICL)", CDPC-AICL(2022)2, 9 June 2022. Available at: https://rm.coe.int/cdpc-aicl-2022-2-2nd-meeting-report/1680a6e1ff.

Council of Europe, "History of Artificial Intelligence". Available at: https://www.coe.int/en/web/artificial-intelligence/history-of-ai.

Council of Europe, Yeung K. , *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*, DGI(2019)05, 2019.

EUROCONTROL, European Aviation/ATM AI High Level Group (EEAI HLG), "Fly AI report", 2020. Available at: https://www.eurocontrol.int/publication/fly-ai-report.

European Commission, Directorate-General for Research and Innovation, Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659), "Ethics of connected and automated vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility", Publications Office of the European Union, Luxembourg, 2020. Available at: <https://data.europa.eu/doi/10.2777/035239>.

European Council on Foreign Relations, U.E. Franke, Policy Brief "Artificial divide: how Europe and America could clash over AI", 2021. Available at: https://ecfr.eu/wp-content/uploads/Artificial-divide-How-Europe-and-America-could-clash-over-AI.pdf.

European Parliament, "How artificial intelligence works", Briefing, Member's Research Service, 2019. Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/634420/EPRS_BRI(2019)634420_EN.pdf.

European Parliament, "Understanding Artificial Intelligence", Briefing, Member's Research Service, 2018. Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/614654/EPRS_BRI(2018)614654_EN.pdf.

European Parliamentary Research Service, Panel for the Future of Science and Technology, "The ethics of artificial intelligence: Issues and initiatives", EPRS_STUD(2020)634452, 2020.

High-Level Expert Group on Artificial Intelligence (AI-HLEG),
A definition of AI: Main capabilities and Scientific Disciplines, 2018. Available at: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf.
Ethics guidelines for trustworthy AI, 2019. Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

OECD, *Recommendation of the Council on Artificial Intelligence,* OECD/LEGAL/0449. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Society for Automotive Engineers International (SAE), *J3016 Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, April 2021.
UNESCO, *Recommendation on the Ethics of Artificial Intelligence,* SHS/BIO/REC-AIETHICS/2021. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

United Nations Economic and Social Council, *Report of the sixty-eighth session of the Working Party on Road Traffic Safety*, ECE/TRANS/WP.1/145, 2014. Available at: https://unece.org/press/unece-paves-way-automated-driving-updating-un-international-convention).

CASE LAW (DOMESTIC DECISIONS)

ITALY
Cassazione Penale, Sez. IV, 25 March 1971.
Cassazione Penale, Sez. IV, 25 March 2016, No. 12748.
GERMANY
BVerfGE, 15.2.2006 - 1 BvR 357/05.
BGH, 18.03.1952, GSSt 2/51.
LG München I, Endurteil vom 14.07.2020 - 33 O 14041/1.
USA
R v M'Naghten's, [1843] All ER Rep 229, 210.
U.S. Supreme Court, Kahler v. Kansas, 23 March 2020, 589 U.S.___ (2020).
Holbrook v. Prodomax Automation Ltd., 1:17-cv-219 (W.D. Mich. Sep. 20, 2021).
United States v. Sun-Diamond Growers of Cal., 138 F.3d 961, 971 (D.C. Cir. 1998), aff'd, 526 U.S. 398, 1999.
Yavapai County Attorney, "Re: Rafael Vasquez / Uber Corporation, Tempe Police Department #2018-32694", 4 March 2019. Available at: https://s3.documentcloud.org/documents/5759641/UberCrashYavapaiRuling03052019.pdf.
State of Arizona, "Indictment 785 GJ 251", 27 August 2020. Available at: http://www.maricopacountyattorney.org/DocumentCenter/View/1724/Rafael-Vasquez-GJ-Indictment.

SECONDARY SOURCES

BOOKS & ARTICLES

Abbott R. & Sarch A., "Punishing Artificial Intelligence: Legal Fiction or Science Fiction", *UCD L Rev*, vo. 53, 2019
Abbott R., "Everything is Obvious", *UCLA Law Review*, Vol. 66, No. 2, 2019.
The Reasonable Robot, Cambridge University Press, 2020.
Abel D., MacGlashan J. & Littman M.L., "Reinforcement Learning as a Framework for Ethical Decision Making", *AAAI Workshop: AI, Ethics, and Society*, 2016.
Allan J., "Revisiting the Hart-Devlin Debate: At the Periphery and By the Numbers", *San Diego L. Rev.*, Vol. 54, 2017, p. 423.
Anscombe G. E. M., "Modern moral philosophy" *Philosophy,* Vol. 33, No. 124, 1968.
Antolisei F., *Manuale di diritto penale-Parte generale*, 16ª ed., Giuffrè, 2003.
Armer P., *Attitudes Toward Intelligent machines,* in Feigenbaum E. A., Feldman J. & Armer P. (Eds.), *Computers and Thought,* Aaai Pr, 1995.

Arnold T. & Scheutz M., "The "big red button" is too late: an alternative model for the ethical evaluation of AI systems", *Ethics and Information Technology*, Vol. 20, 2018.

Asaro P. M.,
"Determinism, machine agency, and responsibility", *Politica & Società,* Vol. 2, 2014.
"A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics", in Lin P., Abney K. & Bekey G. A., *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, 2012.

Ashfarian H., "Can Artificial Intelligences Suffer from Mental Illness? A Philosophical Matter to Consider", *Sci Eng Ethics*, Vol. 23, 2017.

Ashton H., "Definitions of intent suitable for algorithms", *Artif Intell Law*, 2022.

Asimov I., *I, Robot*, Harper Collins Publisher, 1950.

Association International de Droit Pénal (AIDP), L. Picotti, General Report - XXI International Congress of Penal Law on 'Artificial Intelligence and Criminal Justice', International Colloquium of Section I (Criminal Law-general part): "Traditional Criminal Law Categories and AI: Crisis or Palingenesis?", 2022.

Avila Negri S. M. C., "Robots as Legal Persons: Electronic Personhood in Robotics and Artificial Intelligence", *Frontiers in Robotics and AI*, Vol. 8, 2021.

Awad E. et al., "Computational ethics"*, Trends in Cognitive Sciences*, Vol. 26, Issue 5, 2022.

Awad E. et al., "Crowdsourcing Moral Machines", *Communications of the ACM*, Vol. 63, No. 3, 2020.

Awad E. et al., "The Moral Machine experiment", *Nature,* Vol. 563, 2018.

Awad E. et al., "Universals and variations in moral decisions made in 42 countries by 70,000 participants", PNAS, Vol. 117, No. 5, 2020.

Awad et. al, "When Is It Acceptable to Break the Rules? Knowledge Representation of Moral Judgement Based on Empirical Data", arXiv:2201.07763, 2022.

B. Miskolczi & Z. "*Büntetőjogi kérdések az információk korában – Mesterséges intelligencia"*, *Big Data, profilozás. HVG-ORA*C, Budapest, 2018.

Badea C. & Artus G., "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents*"*, in Bramer M. & Stahl F. (Eds.), *Artificial Intelligence XXXIX*. SGAI-AI 2022. *Lecture Notes in Computer Science*, Vol 13652, Springer, 2022.

Baker D. J. & Robinson P. H., *Artificial intelligence and the Law. Cybercrime and Criminal Liability*, Routledge, 2021.

Bandura A., "Moral disengagement in the perpetration of inhumanities*"*, *Personality and Social Psychology Review*, Vol. 3, 1999.

Banks V.A. et al., "Subsystems on the road to full vehicle automation: hands and feet free but not 'mind' free driving", in *Safety Science*, Vol. 62, 2014, pp. 505-514.

Barfield W. & Pagallo U. (Eds.), *Research handbook on the law of artificial intelligence*, Elsevier, 2019.

Basile F., "Intelligenza artificiale e diritto penale: qualche aggiornamento e qualche nuova riflessione", in F. Basile, M. Caterini & S. Romano (Eds.), *Il sistema penale ai confini delle hard sciences*, Pacini Giuridica, 2020.
"Intelligenza artificiale e diritto penale: quattro possibili percorsi di indagine", *DPU*, 2019.

Basile F., Caterini M. & Romano S. (Eds.), *Il sistema penale ai confini delle hard sciences*, Pacini Giuridica, 2020.

Basso P., "*Mare magnum*", *Parolechiave,* Vol. 1, 2022.

Bathaee Y. , "The Artificial Intelligence Black Box and the Failure of Intent and Causation", *Harvard Journal of Law & Technology*, Vol. 31, 2018.

Battaglini G., "The Fascist Reform of the Penal Law in Italy", 24 *Am. Inst. Crim. L. & Criminology* , Vol. 24, 1933-1934.

Beck S.,

"Die Diffusion strafrechtlicher Verantwortlichkeit durch Digitalisierung und Lernende Systeme", *Zeitschrift für Internationale Strafrechtsdogmatik*, Vol. 2, 2020.

"Robotics and Criminal Law. Negligence, Diffusion of Liability and Electronic Personhood", in Hilgendorf E. & Feldle J. (Eds), "Digitization and the Law", *Robotik und Recht*, Vol. 15, Nomos, 2018.

"Intelligent Agents and Criminal Law—Negligence, Diffusion of Liability and electronic personhood", *Robotics and Autonomous Systems*, Vol. 86, 2016.

"Mediating the Different Concepts of Corporate Criminal Liability in England and Germany", *German L.J.*, No. 11, 2010.

Becker G., "Crime and Punishment: An Economic Approach", *J. Pol. Econ,* Vol. 76, 1984.

Ben-Israel I. et al., "Towards Regulation of AI Systems. Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law", Council of Europe Study DGI (2020) 16, 2020.

Bertolini A., "Artificial Intelligence and Civil Liability", Study commissioned by the European Parliament's Committee on Legal Affairs, European Union, IPOL_STU(2020)621926, 2020. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/621926/IPOL_STU(2020)621926_EN.pdf.

Bhargava V. R. & Velasquez M., "Corporate Responsibility And Artificial Intelligence", *The Georgetown Journal of Law & Public Policy*, 2019.

Black H. C., *Black's Law Dictionary,* 4th Ed, West Publishing, 1968.

Blomsma J. and Roef D., "Forms and Aspects of Mens Rea", in Keiler J. & Roef D. (Eds.), *Comparative Concepts of Criminal Law*, 3rd Ed., Intersentia, 2019.

Bohlander M., *Principles of German criminal law*, Hart Publishing, 2009.

Bommarito M. J. II, Katz D. M. & Detterman E. M., "LexNLP: Natural language processing and information extraction for legal and regulatory texts", in Vogl R. (Ed)., *Research Handbook on Big Data Law,* Edward Elgar Publishing, 2021.

Bonicalzi S. & Haggard P., "Responsibility Between Neuroscience and Criminal Law. The Control Component of Criminal Liability", *Rivista internazionale di Filosofia e Psicologia*, Vol. 10, No. 2, 2019.

Borsari R., "Intelligenza Artificiale e responsabilità penale: prime considerazioni", *MediaLaws*, Vol. 3, 2019.

Bostrom N.,

"Ethical Issues in Advanced Artificial Intelligence", 2003. Available at: https://nickbostrom.com/ethics/ai.

"How Long Before Superintelligence?", *International Journal of Futures Studies*, Vol. 2, 1998.

Bovens M., *The quest for responsibility. Accountability and citizenship in complex organizations*, Cambridge University Press, 1998.

Bratman M., *Intention, plans, and Practical Reason*, Harvard University Press, 1987.

Brown D. K., *Public Welfare Offenses*, in Dubbler M. & Hörnle T. (Eds., T*he Oxford Handbook of Criminal Law*, Oxford University Press, 2014.
Brown N. & Sandholm T., "Superhuman AI for heads-up no-limit poker: Libratus beats top professionals", *Science*, Vol. 350, No. 6374, 2017.
Brusco C., "Rischio e pericolo, rischio consentito e principio di precauzione. la c.d. "flessibilizzazione delle categorie del reato", *Criminalia*, 2012.
Bryson J. J., Diamantis M. E. & Grant T. D. , "Of, for, and by the people: the legal lacuna of synthetic Persons", *Artif Intell Law*, Vol. 25, 2017.
Burchard K., "Künstliche Intelligenz als Ende des Strafrechts? Zur algorithmischen Transformation der Gesellschaft", *Normative Orders Working Paper*, No. 2, 2019.
Campbell M., Hoane A.J. Jr. & Hsu F., "Deep Blue", *Artificial Intelligence*, Vol. 134, 2002.
Cappellini A., "Machina delinquere potest? Brevi appunti su intelligenza artificiale e responsabilità penale", *Criminalia*, 2018.
Caruso G., "Free Will Skepticism and Its Implications: An Argument for Optimism", in Shaw E. & Pereboom D. (Eds.)*, Law and Society,* Cambridge University Press, 2019.
Castelvecchi D., "Can we open the black box of AI?", *Nature*, Vol. 538, 5 October 2016.
Cavaceppi C., "L'intelligenza artificiale applicata al diritto penale", in Taddei Elmi G.& Contaldo A., *Intelligenza artificiale-Algoritmi giuridici: Ius condendum o fantadiritto?,* Pacini Giuridica, 2020.
Chalkidis I. et al., "LEGAL-BERT: The Muppets straight out of Law School", *Findings of EMNLP*, 2020.
Chalkidis I., Fergadiotis M. & Androutsopoulos I., "MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer", in Moens M. et al. (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 2021.
Chao V., *Action and Agency in the Criminal Law, Legal Theory*, Cambridge University Press, 2009.
Charisi V., "Towards Moral Autonomous Systems", *arXiv:1703.04741v3*, 2017.
Charney R. , "Can Androids Plead Automatism - A Review of When Robots Kill: Artificial Intelligence under the Criminal Law by Gabriel Hallevy", U. Toronto Fac. L. Rev., Vol. 73 , 2015.
Chesterman S.,
"AI and the Limits of Legal Personality", *International and Comparative Law Quarterly*, Vol. 69, No.4, 2020.
We, the Robots? Regulating Artificial Intelligence and the Limits of the Law, Cambridge University Press, 2021.
Chopra S. & White L. , *A legal theory for autonomous artificial agents*, University of Michigan Press, 2011.
Ciccone J. R. & Ferracuti S., "Comparative Forensic Psychiatry: II. The Perizia and the Role of the Forensic Psychiatrist in the Italian Legal System", *Bull Am Acad Psychiatry Law*, Vol. 23, No. 3, 1995.
Civello G., *La "colpa eventuale" nella società del rischio. Epistemologia dell'incertezza e "verità soggettiva" della colpa*, Giappichelli, 2013.
Claessen J., "Theories of Punishment", in Keiler J. & Roef D. (Eds.), *Comparative Concepts of Criminal Law*, 3rd Ed., Intersentia, 2019.
Coffee J. C. Jr, "'No Soul to Damn: No Body to Kick': An Unscandalized Inquiry into the Problem of Corporate Punishment", *Michigan Law Review*, Vol. 79, No.3, 1981.
Cohen M. K., Hutter M. & Osborne M. A., "Advanced artificial agents intervene in the provision of reward"*, AI Magazine*, Vol. 43, 2022.

Colucci C., *Tra ottimizzazione della funzione comando e prospettive di un suo superamento: i nuovi scenari della normatività penale*, PhD Dissertation, University of Florence Repository, 2022. Available at: https://flore.unifi.it/handle/2158/1273584.

Consulich F., "*Flash offenders*. Le prospettive di *accountability* penale nel contrasto alle intelligenze artificiali devianti*", Rivista Italiana di Diritto e Procedura Penale*, Vol. 3, 2022.

Consulich F., "Le prospettive di *accountability* penale nel contrasto alle intelligenze artificiali devianti", *Diritto e Procedura Penale*, Vol. 3, 2022.

Corsi J. L., "An Argument for Strict Legality in International Criminal Law", *Georgetown Journal of International Law*, Vol. 49, 2018,

D'Addosio C., *Bestie delinquenti*, L. Pierro, 1892

Damgaard Thaysen J., Defining Legal Moralism, *SATS*, Vol. 16, No. 2, 2015.

Danaher J., "Robots, Law and the Retribution Gap", *Ethics and Information Technology*, Vol. 18, No. 4, 2016.

De Maglie C., "Models of Corporate Criminal Liability in Comparative Law", *Wash. U. Global Stud. L. Rev.*, Vol. 4, 2005.

DeArman A., "The Wild, Wild West: A Case Study of Self Driving Vehicle Testing in Arizona", *Arizona Law Review*, Vol. 61, 2019.

DeGrave A. J., Janizek J. D. & Lee S. , "AI for radiographic COVID-19 detection selects shortcuts over signal", *Nature Machine Intelligence*, Vol. 3, 2021.

Dekker S., "*Pilots, Controllers and Mechanics on Trial: Cases, Concerns and Countermeasures"*, *International Journal of Applied Aviation Studies*, Vol. 10, No. 1, 2010.

Diamantis M.E.,

"The Extended Corporate Mind: When Corporations use AI to Break the Law", *N.C. L. Rev.*, Vol. 98, 2020.

"Clockwork Corporations", *Iowa Law Review*, Vol. 1032, 2020.

"Algorithms acting badly: A Solution from Corporate Law", *GEO. Wash. L. Rev.*, Vol. 89, 2021.

"Vicarious Liability for AI", *Cambridge Handbook of AI and Law*, U. *Iowa Legal Studies Research Paper,* No. 2021-27, 2021. Available at SSRN: https://ssrn.com/abstract=3850418,

"Employed Algorithms: a Labor Model of Corporate Liability for AI*", Duke L.J.*, Vol 72, 2022.

Diamantis M. E., Cochran R. & Dam M., "AI and the Law: Can Legal Systems Help Us Maximize Paperclips while Minimizing Deaths?", 2022. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4177378.

Digest of Justinian, 48.19.18, Ulp. 3 ad ed.

Domingos P., *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Allen Lane, 2015.

Doncieuxet S. et al., "Open-Ended Learning: A Conceptual Framework Based on Representational Redescription", in *Frontiers in neurorobotics*, Vol. 12, p. 59, 2018.

Donini M., "An impossible exchange? prove di dialogo tra civil e common lawyers su legalità, morale e teoria del reato"*, Rivista Italiana di Diritto e Procedura Penale*, Fasc.1, 2017.

Donne J., "Meditation XVII. Nunc Lento Sonitu Dicunt, Morieris", in A. Raspa (Ed.), *Devotions Upon Emergent Occasions*, Oxford University Press, 1987.

Dubber M. D. & Hörnle T. (Eds.), *The Oxford Handbook of Criminal Law*, Oxford University Press, 2014.

Dubber M., "The Comparative History and Theory of Corporate Criminal Liability*", New Criminal Law Review: An International and Interdisciplinary Journal* , Vol. 16, No. 2, 2013.

Duff R. A.,
"Moral and Criminal Responsibility: Answering and Refusing to Answer", in J. Coates & N.A. Tognazzini, *Oxford Studies in Agency and Responsibility Volume 5: Themes from the Philosophy of Gary Watson*, 2019.
"Responsibility, Restoration, and Retribution", in M. Torny (Ed.), *Retributivism Has a Past: Has it a Future,* Oxford University Press, 2011.
"Who is Responsible for What, to Whom?", *Ohio State Journal of Criminal Law*, vol. 2, 2005.
"Criminalizing Endangerment", *Louisiana Law Review*, Vol. 65, 2005.
"Legal and Moral Responsibility", *Philosophy Compass* vol. 4 issue 6, 2009.
Intention, Agency and Criminal Liability: Philosophy of Action and the Criminal Law, Blackwell, 1990.
"Towards a Modest Legal Moralism"; *Crim. Law Philos*., Vol. 8, 2014.
Duff, R. A. & Hörnle, T. "Crimes of Endangerment", in Ambos K. et al. (Eds.), *Core Concepts in Criminal Law and Criminal Justice,* Vol. 2, Cambridge University Press, 2022.
Elish M. C. & Hwang T., "Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation", *Comparative Studies in Intelligent Systems – Working Paper #1,* Vol. 2, 2015.
Erling E., "Economics of Criminal Behavior", in B. Bouckaert & G. De Geest, *Encyclopedia of Law & Economics*, Elgar Publishing, 1997.
Evans E. P., *The Criminal Prosecution and Capital Punishment of Animals,* Farber & Farber, 1987.
Fernandes Godinho I., "Law and Science: The Autonomy and Limits of Culpability as a Cornerstone to the Ascription of Liability (or the Subject of Criminal Law: Three Maxims, a Problem and a Glimpse into the Future)", *Int J Semiot Law*, 2022.
Ferzan K. K., "Opaque Recklessness", *Journal of Criminal Law and Criminology*, vol 91, Issue 3, 2001.
Fiandaca G. & Musco E., *Diritto penale. Parte generale*, Zanichelli editore, 7ª ed., 2019.
Fisher T., *Economic Analysis of Criminal Law*, in Dubber M. D. & Hörnle T. (Eds.), *The Oxford Handbook of Criminal Law*, Oxford University Press, 2014.
Fjeld J. et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI", *Berkman Klein Center for Internet & Society*, 2020.
Fletcher G.,
"Criminal Theory in the Twentieth Century", *Theoretical Inquiries in Law,* 2, No. 1, 2001.
"Deutsche Strafrechtsdogmatik aus ausländischer Sicht", in A. Eser, W. Hassemer & B. Burckhardt (Eds.), *Die deutsche Strafrechtswissenschaft vor der Jahrtausendwende,* Beck, 2000.
"The Theory of Criminal Negligence: a Comparative Analysis", *University of Pennsylvania Law Review*, Vol. 119, 1971.
Floridi L., "AI and Its New Winter: from Myths to Realities", in *Philosophy & Technology,* Vol. 33*,* 2020.
Folberth A. et. al, Karlsruhe: Karlsruhe Institute of Technology (KIT), Institute for Technology Assessment and Systems Analysis (ITAS), "Tackling problems, harvesting benefits – A systematic review of the regulatory debate around AI", *KIT Scientific Working Papers*, Vol. 197, 2022.
Foot P., "The Problem of Abortion and the Doctrine of the Double Effect" *Oxford Review*, 1967.

Fosch-Villaronga E. & Heldeweg M., "Regulation, I presume?" said the robot – Towards an iterative regulatory process for robot governance", *Computer Law & Security Review: The International Journal of Technology Law and Practice*, 2018.

Freitas P. M., Andrade F. & Novais P., "Criminal Liability of Autonomous Agents: from the unthinkable to the plausible" in P. Casanovas et al. (Eds.), *AI Approaches to the Complexity of Legal Systems. AICOL 2013 International Workshops, AICOL-IV@IVR, Belo Horizonte, Brazil, July 21-27, 2013 and AICOL-V@SINTELNET-JURIX, Bologna, Italy, December 11, 2013, Revised Selected Papers*, Lecture Notes in Computer Science, Vol. 8929, Springer, 2014.

French P.A., *Collective and Corporate Responsibility*, Columbia University Press, 1984, p. 40

Gaakeer J., "'Sua cuique persona?' A Note on the Fiction of Legal Personhood and a Reflection on Interdisciplinary Consequences", *Law & Literature*, Vol. 28, No. 3, 2016.

Gabriel I., "Artificial Intelligence, Values, and Alignment", *Minds and Machines*, Vol. 30, 2020.

Gaede K., "Künstliche Intelligenz – Rechte und Strafen für Roboter? Plädoyer für eine Regulierung künstlicher Intelligenz jenseits ihrer reinen Anwendung", in E. Hilgendorf & S. Beck (Eds), *Robotik und Recht*, Vol. 18, Nomos, 2018.

Ganascia J. G., "Ethical System Formalization using Non-Monotonic Logics", *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29, 2007.

Gerber M. M. & Jackson J., "Retribution as revenge and retribution as just deserts", *Social Justice Research*, Vol. 26, No. 1, 2013.

Giannini A. & Kwik H. J., "Negligence Failures and Negligence Fixes. A Comparative Analysis of Criminal Regulation of AI And Autonomous Vehicles", Criminal Law Forum, 2023.

Giannini A., "Artificial intelligence, ethics, law: a view on the Italian and American debate (and on their differences)", *Netherlands Journal of Legal Philosophy*, Vol. 51, No. 2, 2022.

Girgen J., "The Historical and Contemporary Prosecution and Punishment of Animals", *Animal Law*, Vol. 9, 2003.

Giuca M., "Disciplinare l'intelligenza artificiale. La riforma francese sulla responsabilità penale da uso di auto a guida autonoma", *Archivio Penale*, Vol. 2, 2022.

Giunta F., "Il diritto penale e le suggestioni del principio di precauzione", *Criminalia*, 2006.

Gless S. & Weigend T. , "Intelligente Agenten und das Strafrecht", *ZSTW*, Vol. 126, No. 3, 2014.

Gless S., Silverman E. & Weigend T., "If Robots Cause Harm, who Is to Blame: Self-Driving Cars and Criminal Liability", *New Criminal Law Review*, Vol. 19, No. 3, 2016,

Goanta C. et al., "Back to the Future: Waves of Legal Scholarship on Artificial Intelligence", in S. Ranchordás & Y. Roznai (Eds.), *Time, Law and Change*, Hart Publishing, 2020.

Gobert J. & Punch M., *Rethinking Corporate Crime*, Cambridge University Press, 2003.

Goldstein A., *The Insanity Defense*, Yale University Press, 1967.

Gomes Rêgo de Almeida P., Denner dos Santos C. & Silva Farias J.," Artificial Intelligence Regulation: a framework for governance", *Ethics and Information Technology*, Vol. 23, 2021.

Goodpaster K. E., "Tenacity: The American Pursuit of Corporate Responsibility", *BUS. & SOC'Y REV.*, Vol. 118, 2013.

Graaf N., *Judicial Influencers: Scholarly use of foreign law and the convergence of German, Italian and French ideas on the position of national constitutional courts in the EU legal context, 1989-2012*, PhD Dissertation, Utrecht University Repository, 2022.

Grandinetti A., "Motivi e movente", *Il Penalista*, Bussola. Available at: https://ilpenalista.it/bussola/motivi-e-movente.

Greco E., *Profili di responsabilità penale del controllore del traffico aereo. Gestione del rischio e imputazione dell'evento per colpa nei sistemi a interazione complessa*, Giappichelli, 2021.
Green B. & and Kak A., The False Comfort of Human Oversight as an Antidote to A.I. Harm Human Agency in Decision-Making Systems", *Slate*, 15 June 2021.
Grosso C. F., Pellissero M. & Petrini D., *Manuale di diritto penale. Parte Generale*, 2ª ed., Giuffrè, 2017.
Guerra A., Parisi F. & Pi D.,
"Liability for robots I: legal challenges", *Journal of Institutional Economics*, Vol 18, Issue 3, 2021.
"Liability for robots II: an economic analysis", Journal of Institutional Economics, Vol. 18, Issue 4, 2021.
Hadzi A. & Roio D., "Restorative Justice in Artificial Intelligence Crime", *Spectres of AI,* Vol. 5, 2019.
Haenlein M. & Kaplan A., "A Brief History of Artificial Intelligence: On the Past, Present and Future of Artificial Intelligence", *California Management Review*, Vol. 61, No. 4, 2019.
Hafter E., *Lehrbuch des Schweizerischen Strafrechts, Allgemeiner Teil*, 2nd Ed., Springer, 1946.
Hall J., *General Principles of Criminal Law, 2nd Ed.*, The Bobbs Merrill Company, 1960.
Hallevy G.,
"The Criminal Liability of Artificial Intelligence Entities - From Science Fiction to Legal Social Control, *Akron Intellectual Property Journal*, Vol. 4, Iss. 2, Art. 1, 2010.
"I, Robot – I, Criminal— When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offences", Syracuse Sci. & Tech. L. Rep., 2010.
"Virtual Criminal Responsibility", 2011. Available at SSRN: https://ssrn.com/abstract=1835362 or http://dx.doi.org/10.2139/ssrn.1835362
"Unmanned Vehicles. Subordination to Criminal Law Under the Modern Concept of Criminal Liability", *Journal of Law, Information and Science,* Vol 21, No. 2002, 2012.
When Robots Kill. Artificial Intelligence under Criminal Law, Northeastern University Press, 2013.
Liability for Crimes Involving Artificial Intelligence Systems, Springer, 2015
"Dangerous Robots – Artificial Intelligence vs. Human Intelligence", 2018. Available at SSRN: https://ssrn.com/abstract=3121905.
"The Basic Models of Criminal Liability of AI Systems and Outer Circles", 2019. Available at SSRN: https://ssrn.com/abstract=3402527 or http://dx.doi.org/10.2139/ssrn.3402527).
Hamlin J. K., "Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core", *Current Directions in Psychological Science,* Vol. 22, Iss. 3, 2013.
Hart H. L. A., *Punishment and Responsibility*, 2nd Ed., Oxford University Press, 2008.
Law, Liberty and Morality, Stanford University Press, 1963.
Hayward K. J., Maas M. M., "Artificial intelligence and crime: A primer for criminologists", *Crime Media Culture*, 2020.
Heaven W.D., "AI is learning how to create itself. Humans have struggled to make truly intelligent machines. Maybe we need to let them get on with it themselves", *MIT Technology Review*, 27 May 2021. Available at: https://www.technologyreview.com/2021/05/27/1025453/artificial-intelligence-learning-create-itself-agi/.

Hildebrandt M.,
Law for Computer Scientists and Other Folk, Oxford University Press, 2020.
"Technology", in Dubber M. D. & Hörnle T. (Eds.), *The Oxford Handbook of Criminal Law*, Oxford University Press, 2014.
"Ambient Intelligence, Criminal Liability and Democracy", *Criminal Law and Philosophy*, Vol. 2, 2008.
Hilgendorf E.,
& J. Feldle (Eds.), *Digitization and the Law*, Nomos, 2018.
"Autonome Systeme, künstliche Intelligenz und Roboter: Eine Orientierung aus strafrechtlicher Perspektive", in Barton S. (Ed.), *Festschrift für Thomas Fischer*, C.H. Beck, 2018.
The dilemma of autonomous driving: Reflections on the moral and legal treatment of automatic collision avoidance systems", in Hilgendorf E. & Feldle J. (Eds), "Digitization and the Law", *Robotik und Recht*, Vol. 15, Nomos, 2018.
"Dilemma-Probleme beim automatisierten Fahren. Ein Beitrag zum Problem des Verrechnungsverbots im Zeitalter der Digitalisierung", *ZIS*, Vol. 130, No. 3, 2018.
Autonome Systeme und neue Mobilität. Ausgewählte Beiträge zur 3. und 4. Würzburger Tagung zum Technikrech, Nomos, 2017.
"Hötitzsch S. & Lutz L. (Eds.), *Rechtliche Aspekte automatisierter Fahrzeuge. Beiträge der 2. Würzburger Tagung zum Technikrecht im Oktober 2014*, Nomos, 2015.
Hofmann R., "Formalism versus pragmatism – A comparative legal and empirical analysis of the German and Dutch criminal justice systems with regard to effectiveness and efficiency", *Maastricht Journal of European and Comparative Law Issue,* Vol. 28, No. 2, 2021.
Holford W. D., "An Ethical Inquiry of the Effect of Cockpit Automation on the Responsibilities of Airline Pilots: Dissonance or Meaningful Control?", *Journal of Business Ethics*, 2022.
Hu Y., "Robot criminals", *U. Mich. J.L. Reform*, Vol. 52, 2019.
Husak D.,
Overcriminalization: The Limits of the Criminal Law, Oxford University Press, 2009.
"The Criminal Law as Last Resort", *Oxford Journal of legal Studies*, Vol. 23, No. 2, 2004, p. 211.
"Theories of Crime and Punishment in German Criminal Law", The American Journal of Comparative Law, 2005, Vol. 53, No. 3, 2005.
Jain N., "Autonomous weapons systems: new frameworks for individual responsibility", in Bhuta N. et al. (Eds), *Autonomous weapons systems. Law, ethics, policy*, Cambridge University Press, 2016.
Jobin A., Ienca M. & Vayena E., "The global landscape of AI ethics guidelines", N*ature Machine Intelligence*, Vol.1, 2019.
K. Nuotio, "Theories of Criminalization and the Limits of Criminal Law: A Legal Cultural Approach", in R.A. Duff et al., The Boundaries of the Criminal Law, Oxford University Press, 2010,
Kalliokoski T., & Hallamaa J., "How AI Systems Challenge the Conditions of Moral Agency?", in Rauterberg M. (Ed.), *Culture and Computing: 8th International Conference, C&C 2020, Held as Part of the 22nd HCI International Conference*, Springer, 2020.
Karsai K., (Ed.), *Strafrechtlicher Lebensschutz in Ungarn und in Deutschland. Beiträge zur Strafrechtsvergleichung*, Stiftung Elemér Pólay, 2007, p. 18.
Keiler J. & Roef D., "Principles of Criminalisation and the Limits of Criminal Law", in Keiler J. & Roef D. (Eds.), *Comparative Concepts of Criminal Law*, 3rd Ed., Intersentia, 2019.
Keiler J., *Actus reus and participation in European Criminal Law,* Intersentia, 2013.

Kim D.J. J., "Artificial intelligence and crime: what killer robots could teach about criminal law", Thesis in Law, Faculty of Law Victoria University of Wellington, 2017. Available at: https://researcharchive.vuw.ac.nz/xmlui/bitstream/handle/10063/7927/paper_access.pdf?sequence=1.

King T. C. et al., "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions", *Sci Eng Ethics,* Vol. 26, No.1, 2019.

Kirchner J. H. et al., "Understanding AI alignment research: A Systematic Analysis", arXiv:2206.02841v1, 2022.

Klip A., *European Criminal Law. An Integrative Approach*, 4th Ed., Intersentia, 2021.

Kneer M. & Stuart M. T., "Playing the Blame Game with Robots", *HRI '21 Companion: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

Koller D & Pfeffer A., "Generating and Solving Imperfect Information Games", P*roceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*,1995.

Korteling J. E. et al., "Human- versus Artificial Intelligence", *Front. Artif. Intell.*, Vol. 4, 2021.

Koulou R., "Human Control over Automation: EU Policy and AI Ethics", *EJLS* 12(1), 2020, pp. 9-46.

Koulou R., "*Human Control over Automation: EU Policy and AI Ethics", EJLS*, Vol. 12, No. 1, 2020.

Kriebitz A., Max R. & Lütge C., "The German Act on Autonomous Driving: Why Ethics Still Matter", *Philosophy & Technology*, Vol. 35, 2022.

Kritikos M., European Parliamentary Research Service – Scientific Foresight Unit, Briefing "Artificial Intelligence ante portas: Legal & ethical reflections", 2019.

Kurki AJ V., A *Theory of Legal Personhood*, Oxford, 2019.

LaCroix, T. & Luccioni A. S., "Metaethical Perspectives on 'Benchmarking' AI Ethicsi", *arXiv:2204.05151*, 2022.

LaFave W. R.,

Criminal Law, 6th Edition, Handbook Series, West Academic, 2017.

Modern Criminal Law: Cases, Comments And Questions, 4th Edition, Thomson West, 1988.

Lagioia F. & Sartor G., "AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective", *Philosophy & Technology*, Vol. 33, 2020.

Langley P., "Explainable, Normative, and Justified Agency"*, The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19).*

Lee S. W., "Can an Artificial Intelligence Commit a Crime?", in Bruns A. et al. (Eds.), *Legal Theory and Interpretation in a Dynamic Society*, Nomos, 2021.

Legg S. & Hutter M., "A Collection of Definitions of Intelligence", *arXiv:0706.3639v1*, 2007.

Lehman-Wilzing S. N., "Frankenstein unbound: Towards a legal definition of artificial intelligence", *Futures*, Vol. 13, Issue 6, 1981.

Leslie D., The Alan Turing Institute, *Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector*, 2019.

Ligeti K., *Artificial Intelligence and Criminal Justice*. Available at: https://www.penal.org/sites/default/files/Concept%20Paper_AI%20and%20Criminal%20Justice_Ligeti.pdf

Lima D., "Could AI Agents Be Held Criminally Liable? Artificial Intelligence and the Challenges for Criminal Law", *South Carolina Law Review*, Vol. 69, Issue 3, 2018.

Lima G. et al., "The Conflict Between People's Urge to Punish AI and Legal Systems", *Front. Robot. AI*, Vol. 8, 2021.

Lin T. et al., "Microsoft COCO: Common Objects in Context, Computer Vision – ECCV", *Lecture Notes in Computer Science*, Springer, Vol. 8693, 2014.

List C. & Petit P., *Group Agency*, Oxford University Press, 2011.

Lo Monte E., "Intelligenza artificiale e diritto penale: le categorie dommatiche alla prova del futuribile", in F. Basile, M. Caterini & S. Romano (Eds.), *Il sistema penale ai confini delle hard sciences*, Pacini Giuridica, 2020.

Luhmann N., *Social Systems*, Stanford University Press, 1995.

Luskin R., "Caring About Corporate 'Due Care': Why Criminal Respondeat Superior Liability Outreaches Its Justification", *American Criminal Law Review*, Vol. 57 2020.

M. Y. Vardi, "Artificial Intelligence: Past and Future", *Communications of the ACM*", Vol. 55, No 1, 2012.

Macagno F., "Definitions in Law", *Bulletin Suisse de Linguistique Appliquee,* 2010.

Magro M. B.,

"Decisione umana e decisione robotica. Un'ipotesi di responsabilità da procreazione robotica", *La legislazione penale*, 2020.

"Biorobotics, robotics and criminal law: some hints and reflections", *Percorsi costituzionali*, Fasc. 1-2, 2016.

"Biorobotica, robotica e diritto penale", in D. Provolo, S. Riondato & F. Yenisey (Eds.), *Genetics, Robotics, Law, Punishment*, 2014.

Malcolm J. G., "Morally Innocent, Legally Guilty: The Case for Mens Rea Reform", *The Federalist Society Review*, Vol. 18, 2017.

Malle B. F. et al, "Which Robot Am I Thinking About? The Impact of Action and Appearance on People's Evaluations of a Moral Robot", *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, 2016.

Malle B. F. et al., "Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents", *HRI '15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction,* 2015.

Malle B. F., Magar S. T. & Scheutz M., "AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma", in Aldinhas M. I. et al. (Eds.), *Robotics and Well-Being. Intelligent Systems*, Control and Automation: Science and Engineering, Vol. 95, Springer, 2019.

Manes V., "L'oracolo algoritmico e la giustizia penale: al bivio tra tecnologia e tecnocrazia", *Discrimen*, 2020.

Mantovani F.,

Diritto penale. Parte speciale. Vol. 1: Delitti contro la persona, 8ª ed., Wolters Kluwer-CEDAM, 2022.

Diritto penale. Parte Generale, 11ª ed, Wolters Kluwer-CEDAM, 2020.

Marcinkevičs R. & Vogt J. E., "Interpretability and Explainability: A Machine Learning Zoo Mini-tour", *ArXiv* abs/2012.01805, 2020.

Marinucci G., Dolcini E. & Gatta G. L., *Manuale di diritto penale. Parte generale*, 8ª ed., Giuffrè, 2019.

Martinho A., "Perspectives about artificial moral agents", *AI and Ethics*, Vol. 1, 2021.

Massi S., "Affidamento sull'intelligenza artificiale e 'disimpegno morale' nella definizione dei presupposti della responsabilità penale", in Giordano R., *Il diritto nell'era digitale. Persona, Mercato, Amministrazione*, *Giustizia*, Giuffrè, 2022.

Matthias A., "The Responsibility Gap", *Ethics and Information Technology,* Vol. 6, No. 3, 2019.

McCarthy J. et al., *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence,* 1955. Available at: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.
McCarthy J., "What is artificial intelligence", 2007. Available at: http://jmc.stanford.edu/articles/whatisai.html.
McGregor S., "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database", *Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21),* 2021.
McGregor S., "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database", *Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21). Virtual Conference*, 2021. Available at https://incidentdatabase.ai/?lang=en.
Miceli T. J., *The Economic Approach to Law*, Stanford University Press, 2004.
Michaelides-Mateou S. & Mateou A., *Flying in the Face of Criminalization. The Safety Implications of Prosecuting Aviation Professionals for Accidents,* Routledge, 2010.
Mill J. S., *On Liberty*, Penguin Books, 2010.
Minkkinen P., "'If Taken in Earnest': Criminal Law Doctrine and the Last Resort", *The Howard Journal*, Vol. 45, No. 5, 2006.
Mokhtarian E., "The Bot Legal Code: Developing a Legally Compliant Artificial Intelligence", *Vanderbilt Journal of Entertainment and Technology Law*, Vol. 21, 2020.
Moore M., *Placing Blame: a General Theory of the Criminal Law*, Clarendon Press, 1997.
Moravec H., *Mind Children. The future of Robot and Human Intelligence*, Harvard University Press, 1990.
Morse S. J., "Rationality and Responsibility", *Southern California Law Review*, Vol. 74, 2000.
Mosier K. & Skitka L.J., "Automation use and automation bias", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1999.
Mulligan C., "Revenge Against Robots", *South Carolina Law Review*, Vol. 69, 2018.
Nalla M., "Singapore", *World Factbook of Criminal Justice Systems*, 1993. Available at: https://bjs.ojp.gov/content/pub/pdf/wfbcjss.pdf.
National Alliance On Mental Illness, "A Guide to Mental Illness and the Criminal Justice System", 2022. Available at: www.nami.org.
Nicholson Price W. II, "Big Data, Patents, and the Future of Medicine", *Cardozo Law Review,* Vol. 37, 2016
Norvig P. & Russel S., *Artificial Intelligence. A Modern Approach*, Pearson Education, 2003.
Osmani N., "The Complexity of Criminal Liability of AI Systems", *Masaryk University Journal of Law and Technology,* Issue 14, No. 1, 2020.
Packard N. et al., "An Overview of Open-Ended Evolution: Editorial Introduction to the Open-Ended Evolution II Special Issue", *Artificial life*, Vol 25, No. 2, 2019.
Pagallo U.,
The Laws of Robots. Crimes, Contracts, and Torts, Springer, 2013.
"The Adventures of Picciotto Roboto", *The Social Impact of Social Computing - Ethicomp*, 2011.
"Killers, Fridges, and Slaves", *AI and Society*, Vol. 26, No. 4, 2011.
"AI and bad robots" in McGuire M. R. & Holt T. (Eds.), *The Routledge Handbook of Technology, Crime and Justice,* Routledge, 2017.
"From automation to autonomous systems: A legal phenomenology with problems of accountability", *IJCAI International Joint Conference on Artificial Intelligence*, Conference paper, 2017.

Pagallo U. & Barfield W., A*dvanced Introduction to Law and Artificial Intelligence*, Edward Elgar Publishing, 2020.
Pagallo U. & Quattrocolo S., "The impact of AI on criminal law, and its twofold procedures", in Barfield W. & Pagallo U. (Eds.), *Research handbook on the law of artificial intelligence*, Elsevier, 2019.
Pagliaro A., *Principi di diritto penale. Parte generale*, 9ª ed., Giuffrè,2020.
Palazzo F., *Corso di diritto penale. Parte generale,* 8ª Ed., Giappichelli, 2021.
Palazzo F.C. & M. Papa, *Lezioni di diritto penale comparato*, 3ª Ed., Giappichelli, 2013.
Panattoni B., "Intelligenza artificiale: le sfide per il diritto penale nel passaggio dall'automazione tecnologica all'autonomia artificiale", *Dir. Inf.*, Vol. 2, 2021.
Papa M., "The Offense Definition as a Screenplay of Evil: The Rise and Fall of Visual Criminal Law", *Católica Law Review*, Vol. 4, No. 3, 2020.
Papa M., *Fantastic Voyage,* 2ndEd., Giappichelli, 2019.
Parasuraman R. & Manzey D. H., "Complacency and Bias in Human Use of Automation: An Attentional Integration", *Human Factors,* Volume 52, No. 3, 2010.
Pare G.et al., "Synthesizing information systems knowledge: A typology of literature reviews", *Information & Management*, No. 52, 2015, p. 183.
Partridge D., *A New Guide to Artificial Intelligence,* Intellect L & D E F A E, 1991, p. 1.
Pattinson J., Haibo C. & Subhajit B., "Legal issues in automated vehicles: critically considering the potential role of consent and interactive digital interfaces", *Humanities and Social Sciences Communications*, Vol. 7, Art. No. 153, 2020.
Peikert A. (Dipl.-Jur.), Reinelt A. (Dipl.-Jur.) & Witt (Dipl.-Jur.) J. , "Diskussionsbeiträge der 38. Tagung der deutschsprachigen Strafrechtslehrerinnen und Strafrechtslehrer 2019 in Hannover", *ZSTW*, Vol. 131, No. 4, 2019.
Peršak N., *The Harm Principle, its Limits and Continental Counterparts*, Springer, 2007.
Picotti L., AIDP Section I "Traditional Criminal Law Categories and AI: Crisis or Palingenesis?", Questionnaire. Available at: https://www.penal.org/sites/default/files/Questionnaires%20EN.pdf.
Piergallini C., "Intelligenza artificiale: da 'mezzo' ad 'autore' del reato?", *Rivista italiana di diritto e procedura penale*, Vol. 4, 2020.
Pizzi M., Romanoff M. & Engelhardt T., "AI for humanitarian action: Human rights and ethics", *International Review of the Red Cross*, Vol. 102, No. 913, 2020.
Poel I.R. et al. (Eds.), *Moral Responsibility and the Problem of Many Hands*, Routledge, 2015
Posner R. A., "An Economic Theory of the Criminal Law", *Columbia Law Review*, Vol. 85, 1985.
Quarck L., "Zur Strafbarkeit von e-Personen", *Zeitschrift für Internationale Strafrechtsdogmatik*, Vol. 2, 2020.
Rengier R., *Strafrecht Allgemeiner Teil*, C.H. Beck, 2019.
Rességuier A. & Rodrigues R., "AI ethics should not remain toothless! A call to bring back the teeth of ethics" *Big Data & Society*, 2020.
Rich E. & Knight K., *Artificial Intelligenc*e, Tata McGraw, 2004.
Riondato S., "Robot: talune implicazioni di diritto penale" in P. Moro & C. Sarra (Eds.), *Tecnodiritto. Temi e informatica e robotica giuridica*, Franco Angeli, 2017.
Rizzo Minelli G., "Quando l'autore del reato è un robot: tra vecchi modelli imputativi e nuovi possibili paradigmi di responsabilità" in F. Basile, M. Caterini & S. Romano (Eds.), *Il sistema penale ai confini delle hard sciences*, Pacini Giuridica, 2020.
Robson R. A., "Crime and Punishment: Rehabilitating Retribution as a Justification for Organizational Criminal Liability", *Am. Bus. L.J.*, Vol. 47, 2010.

Rodotà S., "Etica e Diritto (dialogo tra alcuni studenti e Stefano Rodotà) con una Presentazione di Gaetano Azzariti", *Costituzionalismo.it,* Vol. 1, 2019.
Rosca C. et al., "Return of the AI: An Analysis of Legal Research on Artificial Intelligence Using Topic Modelling", *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop*, 2020.
Rosenberg M.T., "The Continued Relevance of the Irrelevance-of-Motive-Maxim", *Duke Law Journal*, Vol. 57, No. 4, 2008.
Rudschies C., Schneider I. & Simon J., "Value Pluralism in the AI Ethics Debate – Different Actors, Different Priorities", *The International Review of Information Ethics*, Vol. 29, 2021.
Russell S. J. & Norvig P., *Artificial Intelligence. A Modern Approach*, 2nd edition, Pearson, 2003.
Salvadori I., "Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale", *Rivista Italiana di Diritto e Procedura Penale*, No. 1, 2021.
Samoili S. et al., "AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence", EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, JRC118163.
Santoni de Sio F., & Mecacci G., Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them", *Philosophy & Technology,* Vol. 34, 2021, p.
Sarch A., "Should Criminal Law Mirror Moral Blameworthiness or Criminal Culpability? A Reply to Husak", *Law and Philosophy*, Vol. 41, 2022.
Sarch A., "Who cares what you think? Criminal Culpability and the Irrelevance of Unmanifested Mental States", *Law and Philosophy*, Vol. 36, No. 6, 2017.
Sartor G., "Decisioni algoritmiche tra etica e diritto", in Ruffolo U. (Ed.), *Intelligenza artificiale. Il diritto, i diritti, l'etica*, Giuffré, 2020.
Scaroina E., "La responsabilità penale del datore di lavoro nelle organizzazioni complesse", *Sistema penale*, 2021.
Scherer M.U., "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies", *Harvard Journal of Law & Technology*, Vol. 29, No. 2, 2016.
Schramowski P. et al., "The Moral Choice Machine", *Frontiers in Artificial Intelligence*, Vol. 3, 2020.
Schuett J., "A Legal Definition of AI", *Xiv:1909.01095 [cs.CY]*, 2019.
Schünemann B., "Über Strafrecht im demokratischen Rechtsstaat, das unverzichtbare Rationalitätsniveau seiner Dogmatik und die vorgeblich progressive Rückschrittspropaganda", *ZIS*, Vol. 10, 2016.
Seher G., "Intelligente Agenten als 'Personen' im Strafrecht?", in Gless S. & Seelmann K. (Eds.), *Intelligente Agenten und das Recht*, Vol. 9, 2016, Nomos.
Serafimova S., "Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement", *Humanities and Social Sciences Communication*, 2020
Shalchi A., House of Commons Library, Research Briefing, "Corporate criminal liability in England and Wales", CBP 9027, 2022
Sheikh H., Prins C., & Schrijvers E., *Mission AI. Research for Policy,* Springer, 2023.
Shenkman C., D. Thakur D. & Llansó E., Center for Democracy & Technology, *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis*, 2021. Available at: https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf.
Shin Y., "The Spring of Artificial Intelligence in Its Global Winter", *IEEE Annals of the History of Computing,* Vol. 41, Issue 4, 2019.

Sikdokur I., Baytas I. & Yurdakul A., "Image Classification on Accelerated Neural Networks", *arXiv:2203.11081v1* [cs.CV], 2022.
Simester A. P., *Fundamentals of Criminal Law: Responsibility, Culpability, and Wrongdoing*, Oxford University Press, 2021.
Simmler M. & Markwalder N., "Guilty Robots? – Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence", *Criminal Law Forum*, No. 30, 2019.
Simmler M., "Automation", in Caeiro P. et al. (Eds.), *Elgar Encyclopedia of Crime and Criminal Justice*, Vol. 1, 2023.
Simon H. A., Shaw J. & Newell A., "Heuristic Problem Solving: The Next Advance in Operations Research", *Operations Research*, Vol. 6, No. 1, 1958.
Simon H. A., Shaw J. & Newell A., "Report on a General Problem-Solving Program", *Rand Corporation*, 1959.
Simpson A. W. B., *Legal Theory and Legal History: Essays on the Common Law*, The Hambledon Press, 1987.
Sing J. L. S., "The Continuing Confusion Over Section 304A of the Singapore Penal Code", *Singapore Journal of Legal Studies*, 2015.
Singapore Penal Code, 1871.
Snyder H., "Literature review as a research methodology: An overview and guidelines", *Journal of Business Research*, Vol. 104(C), 2019.
Søbirk Petersen T., "What is Legal Moralism?", *SATS*, Vol. 12, 2011.
Solum L., "Legal Personhood for Artificial Intelligences", *N.C. L. Rev.*, Vol. 7, 1992.
Srinivasa S. & Deshmukh J., "AI and the Sense of Self", *ArXiv* abs/2201.05576, 2021.
Stanley Kenneth O., "Why Open-Endedness Matters", *Artif Life*, Vol. 25, No. 3, 2019.
Stone P. et al., "Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel", Stanford University, September 2016. Available at: http://ai100.stanford.edu/2016-report.
Stonier T., *Beyond Information. The Natural History of Intelligence*, Springer, 1992.
Storrs Hall J., "Towards Machine Agency: a Philosophical and Technological Roadmap", 2012. Available at: https://robots.law.miami.edu/wp-content/uploads/2012/01/Hall-MachineAgencyLong.pdf>.
Sultana F., Sufian A. & Dutta P., "Advancements in Image Classification using Convolutional Neural Network", *arXiv:1905.03288v1* [cs.CV], 2019.
Surden H., "Artificial Intelligence and Law: an Overview", *Ga. St. U. L. Rev.*, Vol. 35, 2019.
T. Padovani, *Diritto penale*, 8ª Ed., Giuffrè, 2019.
Talamo V.C., "Sistemi di intelligenza artificiale: quali scenari in sede di accertamento della responsabilità penale?", *Il Penalista*, 2020.
Thompson D. F., "Designing Responsibility: The Problem of Many Hands in Complex Organizations", *The American Political Science Review,* Vol. 74, No. 4, 1980,
Thomson J. J. "The Trolley Problem" *The Yale Law Journal,* Vol. 94, No.6, 1985.
Tigard D. W., "There Is No Techno-Responsibility Gap", *Philosophy & Technology*, Vol. 1, 2021.
Tsarapatsanis D. & Aletras N., "On the Ethical Limits of Natural Language Processing on Legal Text", *ACL-IJCNLP*, 2021.
Turing A., "Computing Machinery and Intelligence", *Mind*, LIX/236, 1950.
Turner J., *Robot Rules. Regulating Artificial Intelligence*, Gildan Media Corporation, 2019.
Tuttle E., "Reexamining the Vicarious Criminal Liability of Corporations for the Willful Crimes of Their Employees", *Clev. St. L. Rev.,* Vol. 70, 2021.

Van de Poel I., Royakkers L. & Zwart S. D. , *Moral Responsibility and the Problem of Many Hands*, Routledge, 2018.
Veneziani P., *Motivi e colpevolezza*, Giappichelli, 2000.
Verdicchio M. & Perin A., "When Doctors and AI Interact: on Human Responsibility for Artificial Risks", *Philosophy & Technology*, Vol. 35, 2022.
Vogel J., "Strafrecht und Strafrechtswissenschaft im internationalen und europäischen Rechtsraum", *ZIS*, No.1, 2012.
Wagner B., "Human Agency in Decision-Making Systems", *Policy & Internet,* Vol. 11, No.1, 2019.
"Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems", *Policy & Internet*, Vol. 11, Issue 1, 2019.
Wallach I., Dzamba M. & Heifets A., "AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery", *ArXiv abs/1510.02855*, 2015.
Walsh C.G. et al., "Prospective Validation of an Electronic Health Record–Based, Real-Time Suicide Risk Model", *JAMA Netw Open*, Vol. 4, No. 3, 2021.
Walsh C.G. & Ribeiro J. D. & Franklin J. C., "Predicting Risk of Suicide Attempts Over Time Through Machine Learning", *Clinical Psychological Science*, Vol.5, Issue 3, 2017.
Waltermann A.,
"Why non-human agency", in A. Waltermann et. al (Eds.), *Law, Science, Rationality*, Maastricht Law Series, No. 14, Eleven International Publishing, 2020.
"On the legal responsibility of artificially intelligent agents: addressing three misconceptions", *Technology and Regulation*, 2021.
Wang P., "On Defining Artificial Intelligence", *Journal of Artificial General Intelligence*, Vol. 10, No. 2, 2019.
"What Do You Mean by 'AI'?", *Proceedings of the 2008 conference on Artificial General Intelligence*, 2008.
Wang R. et al., "Paired Open-Ended Trailblazer (POET): EndlesslyGenerating Increasingly Complex and Diverse Learning Environments and Their Solutions", *ArXiv abs/1901.01753*, 2019.
Wang Y. & Kosinski M., "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images", *Journal of Personality and Social Psychology*, Vol. 114, Issue 2, 2018, pp. 246–257.
Weigend T., "Subjective Elements of Criminal Liability", in Dubber M. D. & Hörnle T. (Eds.), *The Oxford Handbook of Criminal Law*, Oxford University Press, 2014
Weizenbaum J., "ELIZA—a computer program for the study of natural language communication between man and machine", *Communications of the ACM*, Vol. 9, Issue 1, 1966.
Wiener E.L., "Complacency: Is the term useful for air safety?", *Proceedings of the 26th Corporate Aviation Safety Seminar*, 1981.
Wu X. et al., "A Survey of Human-in-the-loop for Machine Learning", *arXiv:2108.00941*, 2021.
Wyde W. W., *The Criminal Prosecution and Capital Punishment of Animals,* W. Heinemann, 1906.
Yeo S., Morgan N. & Wing Cheong C., *Criminal Law in Malaysia and Singapore,* 2nd Ed., LexisNexis, 2012.
Yu H., "From Deep Blue to DeepMind: What AlphaGo Tells Us", *Predictive Analytics and Futurism*, Issue 13, 2016.

Yudkwosky E., "The AI Alignment Problem: Why It's Hard, and Where to Start", Transcription of the speech given at Stanford University on May 5, 2016. Available at: https://intelligence.org/files/AlignmentHardStart.pdf.
Zanzotto F. M., "Human-in-the-loop Artificial Intelligence", *Journal of Artificial Intelligence Research*, Vol. 64, 2019.
Zhong H. et al., "How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
Zhou Q. et al., "DecisionHoldem: Safe Depth-Limited Solving With Diverse Opponents for Imperfect-Information Games", *ArXiv:2201.11580*, 2022.
Smiley L., "'I am the Operator': The Aftermath of a Self-Driving Tragedy", *WIRED*, 8 March 2022. Available at: https://www.wired.com/story/uber-self-driving-car-fatal-crash/.
Nichols G., "Surgery digitized: Telesurgery becoming a reality", *ZD Net,* 14 June 2021. Available at: https://www.zdnet.com/article/surgery-digitized-telesurgery-becoming-a-reality/.

NEWSPAPER ARTICLES & BLOG POSTS

"What is machine learning data poisoning?", *TechTalks*, 7 October 2020. Available at: https://bdtechtalks.com/2020/10/07/machine-learning-data-poisoning/. English R., "From tiger to free-range parents – what research says about pros and cons of popular parenting styles", *The Conversation*, 25 May 2016. Available at: https://theconversation.com/from-tiger-to-free-range-parents-what-research-says-about-pros-and-cons-of-popular-parenting-styles-57986.
BBC, "Uber's self-driving operator charged over fatal crash", 16 September 2020,. Available at: https://www.bbc.com/news/technology-54175359.
Casey K., "How to explain machine learning in plain English", *The Enterprisers Project*, 19 November 2020. Available at: https://enterprisersproject.com/article/2019/7/machine-learning-explained-plain-english?page=0%2C0.
Dao D. et al., "Awful AI - 2021 Edition". Available at: https://github.com/daviddao/awful-ai.
Darrach B., "Meet Shaky, the first electronic person", *Life Magazine*, 20 November 1970. Available at: https://books.google.it/books?id=2FMEAAAAMBAJ&lpg=PA57&dq=%22first%20electronic%20person%22&pg=PA58#v=onepage&q=years&f=false.
Dickson B., "What are artificial neural networks (ANN)?", *TechTalks*, 5 August 2019. Available at: https://bdtechtalks.com/2019/08/05/what-is-artificial-neural-network-ann/.
Ezra Klein Interviews Alison Gopnik, The New York Times, 16 April 2021. Transcript available at: https://www.nytimes.com/2021/04/16/podcasts/ezra-klein-podcast-alison-gopnik-transcript.html.
Floridi L., Should we be afraid of AI?, *aeon.co*. Available at: https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible.
Goldhill O., "Machines know when someone's about to attempt suicide. How should we use that information?", *Quartz*, 5 September 2018. Available at: https://qz.com/1367197/machines-know-when-someones-about-to-attempt-suicide-how-should-we-use-that-information/.

Graaf N., "Why German Law Libraries Are Not Neutral and Why We Should Care", *LawLog* Blog, July 2019. Available at: https://lawlog.blog.wzb.eu/2019/07/25/why-german-law-libraries-are-not-neutral-and-why-we-should-care/.

Grant K., "Random darknet shopper exhibition featuring automated dark web purchases opens in London", *The Independent*, 2 December 2015. Available at: https://www.independent.co.uk/life-style/gadgets-and-tech/news/random-darknet-shopper-exhibition-featuring-automated-dark-web-purchases-opens-in-london-a6770316.html.

Hunt E., "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter", *The Guardian*, 24 March 2016. Available at: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

Jahromi N., "The Unexpected Philosophical Depths of the Clicker Game Universal Paperclips", *The New Yorker online*, 28 March 2019. Available at: https://www.newyorker.com/culture/culture-desk/the-unexpected-philosophical-depths-of-the-clicker-game-universal-paperclips.

Kang D. et al., "Finding Errors in Perception Data With Learned Observation Assertions", *Stanford Dawn*, 24 January 2022. Available at: https://dawn.cs.stanford.edu/2022/01/24/loa/.

Knight W., "Two rival AI approaches combine to let machines learn about the world like a child", *MIT Technology Review*, 8 April 2019. Available at: https://www.technologyreview.com/2019/04/08/103223/two-rival-ai-approaches-combine-to-let-machines-learn-about-the-world-like-a-child/.

Krisher T. & Dazio S., "Felony Charges Are 1st in a Fatal Crash Involving Autopilot" *AP NEWS,* Los Angeles, 18 January 2022. Available at: https://apnews.com/article/tesla-autopilot-fatal-crash-charges-91b4a0341e07244f3f03051b5c2462ae.

Martineau K., "Teaching machines to reason about what they see", *MIT News*, 2 April 2019. Available at: https://news.mit.edu/2019/teaching-machines-to-reason-about-what-they-see-0402.

Power M., "What happens when a software bot goes on a Darknet shopping spree?", *The Guardian*, 5 December 2014. Available at: www.theguardian.com/technology/2014/dec/05/software-bot-darknet-shopping-spree-random-shopper.

Press G., "The Brute Force of IBM Deep Blue And Google DeepMind", *Forbes online*, 7 February 2018. Available at: https://www.forbes.com/sites/gilpress/2018/02/07/the-brute-force-of-deep-blue-and-deep-learning/?sh=597fb30249e3.

Quirk C., "So Smart It's Stupid", *Alumni*, 1 April 2022. Available at: https://alumni.msu.edu/stay-informed/alumni-stories/so-smart-its-stupid-the-weirdness-of-ai.

Reuters, "Dutch vehicle authority seeks answers on fatal Tesla crash", 14 July 2016. Available at: https://www.reuters.com/article/tesla-authority-dutch-idINL8N1A03KF.

Rogers A., "The Way the World Ends: Not with a Bang But a Paperclip", *Wired*, 21 October 2017. Available at https://www.wired.com/story/the-way-the-world-ends-not-with-a-bang-but-a-paperclip/.

Solon O., "Oh the humanity! Poker computer trounces humans in big step for AI", *The Guardian*, 31 January 2017. Avaialble at:

https://www.theguardian.com/technology/2017/jan/30/libratus-poker-artificial-intelligence-professional-human-players-competition.
Spector M. & Levine D., "Exclusive: Tesla faces U.S. criminal probe over self-driving claims" *Reuters,* 27 October 2022. Available at: www.reuters.com/legal/exclusive-tesla-faces-us-criminal-probe-over-self-driving-claims-sources-2022-10-26/.
Suwandi R. C., "Why is AI So Smart and Yet So Dumb? What Moravec's Paradox told us about AI", *Towards data science*, 30 august 2020. Available at: https://towardsdatascience.com/why-ai-is-so-smart-and-yet-so-dumb-c156cc87fafa.
The Guardian, "Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian", 19 March 2018. Available at: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.
Wakabayashi D., "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam", *The New York Times online*, 19 March 2018. Available at: https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.
Zewe A., "Avoiding shortcut solutions in artificial intelligence', *MIT news*, 2 November 2021. Available at: https://news.mit.edu/2021/shortcut-artificial-intelligence-1102.

OTHER SOURCES

!Mediengruppe Bitnik & !Digital Brainstoarming, "The Darknet – From Memes to Onionland. An Exploration", Kunst Halle Sankt Gallen.
!Mediengruppe Bitnik, Random Darknet Shopper, 2014-2016. Available at: https://wwwwwwwwwwwwwwwwwwwwww.bitnik.org/r/.
"Paperclip maximizer", *Arbital*. Available at: https://arbital.com/p/paperclip_maximizer/
"ResponsibleAI". Available at: https://romanlutz.github.io/ResponsibleAI/.
"Squiggle Maximizer (formerly "Paperclip maximizer"). Available at: https://www.lesswrong.com/tag/paperclip-maximizer.
"Utility functions". Available at: https://www.lesswrong.com/tag/utility-functions.
AI Incident database: https://incidentdatabase.ai/.
AIAAIC Repository: https://docs.google.com/spreadsheets/d/1Bn55B4xz21-_Rgdr8BBb2lt0n_4rzLGxFADMlVW0PYI/edit#gid=1051812323.
AIAAIC Repository. Available at: https://docs.google.com/spreadsheets/d/1Bn55B4xz21-_Rgdr8BBb2lt0n_4rzLGxFADMlVW0PYI/edit#gid=1051812323.
AIDP, XXIst International Congress of Penal Law, 2024: http://www.penal.org/en/information
Airbus, "Airbus concludes ATTOL with fully autonomous flight tests", 29 June 2020. Available at: https://www.airbus.com/en/newsroom/press-releases/2020-06-airbus-concludes-attol-with-fully-autonomous-flight-tests.
Cassandra R. A., "The POMDP Page". Available at: https://www.pomdp.org/.
ChatGPT: chat.openai.com.
DataCentricAI: https://datacentricai.org.
IBM, "Deep Blue". Available at: https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/.
Investopedia, "Feasibility Study", 4 November 2022. Available at: https://www.investopedia.com/terms/f/feasibility-study.asp.
Kavlakoglu E., "AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?", *IBM*, 27 May 2020. Available at:

https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks.
Lackman J., Random Darknet Shopper, *Aksioma – Institute for Contemporary Art,* 2016. Available at: https://aksioma.org/pdf/aksioma_PostScriptUM_23_ENG_Bitnik.pdf.
LessWrong,
National Highway Traffic Safety Administration, "Automated Vehicles for Safety". Available at: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety.
NormanAI: http://norman-ai.mit.edu/.
The MIT's Moral Machine: https://www.moralmachine.net/
The Paperclip Maximizer: https://www.decisionproblem.com/paperclips/index2.html.

BIOGRAPHY

Alice Giannini was born in Castelfranco Veneto (TV), Italy, on August 6, 1993. She was a PhD candidate in Criminal Law at the University of Florence and at Maastricht University from 2019 to 2023.

In 2021 she won the first edition of the Giulia Cavallone price to finance her research stay at Maastricht University. The price was awarded by the Foundation "Centro di Iniziativa Giuridica Piero Calamandrei" and the Cavallone family to honor the memory of Dr. Giulia Cavallone, judge of the Court of Rome and PhD in criminal law and procedure.

During her PhD, she gained extensive teaching experience in the field of international criminal law and comparative criminal law, both in Italian and in English, and became a member of the Law and Neurosciences research network (LACS), AI & Neurotechnology cluster.

She became a lawyer in November 2019 and has a combined bachelor and master degree in law (*Laurea magistrale in Giurisprudenza*) from the University of Bologna – Alma mater studiorum.

Over the years, she has been an active member of several non-profit organizations which provided pro-bono legal counsel to homeless people and prisoners. She is currently a member of StraLi for Strategic Litigation, an Italian NGO which pursues strategic litigation to further the protection of human rights at national, and supranational, level.