UiO : Faculty of Law
University of Oslo

# Disentangling criminal accountability for robot harms

An alternative view to the problem of attributing of criminal liability for the harms of robotic and AI systems

Candidate number: 9001

Submission deadline:30.09.2021

Number of words: 39,972

# Table of contents

# Table of abbreviations

| Abbreviation | Definition |
|---|---|
| ADS | Automatic Driving System |
| ACHPR | African Charter on Human and Peoples' Rights |
| ACHR | American Convention on Human Rights |
| AComHPR | African Commission on Human and Peoples' Rights |
| ACtHPR | African Court on Human and Peoples' Rights |
| AI | Artificial intelligence |
| CAT | Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment |
| CAV | Connected and autonomous vehicle |
| CPPCG | Convention on the Prevention and Punishment of the Crime of Genocide |
| CPDC | European Committee on Crime Problems |
| CPS | Cyber-physical system |
| (The) Convention | European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14 |
| ECtHR/the Court/the Strasbourg Court (or ECHR in footnotes) | European Court of Human Rights |
| EC | European Commission |
| EU | European Union |
| EP | European Parliament |
| GCIV | Geneva Convention Relative to the Protection of Civilian Persons in Time of War |
| HRW | Human Rights Watch |
| ICHR | Inter-American Commission of Human Rights |
| ICCPR | International Covenant on Civil and Political Rights |
| ICtHR | Inter-American Court of Human Rights |
| IHL | International Humanitarian Law |
| IHRC | International Human Rights Clinic at Harvard Law School |
| ML | Machine learning |
| UNGA | United Nations General Assembly |
| UN-SC | United Nations Security Council |

# 1 Introduction

> *A culture of accountability is particularly important for a technology still struggling with standards of reliability because it means that even in cases where things go awry, we are assured of answerability*[1]

## A. Intelligent artefacts and the diffusion of criminal accountability: subject matter

A feature of current times is that machines are taking over more and more decisions. Typical cars would slavishly respond to the commands of humans steering the wheel, whereas emerging connected and autonomous vehicles (CAV(s) increasingly decide when to brake or change the lane. And those decisions go as far as to include life-and-death choices. Like other robots, these vehicles will have to choose between, among others, protecting their passengers or avoiding a collision with pedestrians.[2] The lethal autonomous weapon system (LAWS) is also a case in point: far from a gun in the hands of a human, the machine decides who is its target and whether to fire or not.[3] Decisions, even those involving most cherished values like life, are in the hands of a machine.

How is that possible? What kind of event is turning artefacts like cars and weapons into decision-makers? On the one hand, developments in computing power coupled with terabytes of data render increasingly intelligent machines. Artificial intelligence (AI) is the field of knowledge behind artefacts that not only emulate human thinking processes but even outsmart them. On the other, there are developments in actuators and sensors. If AI produces a brain, these later developments give the machine senses and arms to both perceive and act upon the world.[4]

These technologies bring tremendous benefits. Robotic wheelchairs that can learn how to navigate complex environments without human control open a world of possibilities for disabled persons.[5] AI might detect signs of lung cancer earlier and faster than trained radiologists,[6] whereas autonomous surgeons promise to alternate with specialists located far from the operating room.[7] And CAVs have the potential of making highways considerably safer than they

[1] Nissenbaum, (1996), 2.

[2] See, e.g., Santoni De Sio, (2017).

[3] See Royakkers and Est, (2015), 559-63.

[4] See Section G.iii.

[5] See, e.g., Simpson et al., (2004); Crisman et al., (1998); Bien et al., (2003).

[6] See, e.g., Sathykumar et al., (2020); Borrelli et al., (2021).

[7] See, e.g., Schrempf and Anthuber, (2019).

are.[8] However, these developments generate significant challenges for the realisation of human rights.

Decision-making comes with expectations of accountability, and as machines replace humans in the former, it is unclear who will assume the latter. When things go wrong, victims and society want to know whether the harm was the result of someone's behaviour and if it is one for which there is no justification or excuse to have its author condemned or even punished.[9] And nothing speaks better of someone's behaviour than her decisions. Traditionally, pointing to the human who made the decision would suffice to answer the question "who did it?" Those practices are nonetheless put to stress insofar as those choices are more the machine's and less of a human agent. If asked "who did it?" a robotic wheelchair user, or a doctor alternating with a robotic surgeon, might answer: the machine did it. But how to punish or even condemn a robot? If not the robot, then who? It seems that the cost to pay for novel technologies and their benefits is the erosion of accountability.

That cost is nowhere else more prominent than in criminal law. Concerned with the most invasive responses to the gravest forms of damage **—**like severe injuries or loss of life**—**, its primary function is to express disapproval with punishment.[10] With its sting, criminal law also seeks to dissuade the blamed person and others from further offending. To perform in such a manner, it is not fitting to identify who is better positioned to pay for the damages or take insurance. It would be unreasonable, for instance, to tell an insurance taker to avoid deeds she did not perform in the first place. Determining "who did it" is, thus, inescapable for criminal law. Yet, that is precisely the element that robotics and smart artefacts threaten.

Imagine a world where machines kill or harm with no one to account for it. What would be the fate of human rights to life, or the prohibition of torture and ill-treatment, without the possibility to sanction and prevent recurrent violations? How to ensure redress and protection of potential victims who are affected by the decisions of these technologies? What would be the fate of accountability as it deals with technologies that —as indicated in the Chapter's epigraph— are still struggling with standards of reliability?

These are not futuristic challenges anymore. "Have autonomous robots started killing in war?" headlines an article[11] echoing an UN-SC report that narrates how drones autonomously "hunted down and remotely engaged" members of the Haftar Affiliated Forces.[12] "Why Was-

[8] See Zimmer, (2017).

[9] Santoni de Sio and Mecacci, (2021), section 2.1.

[10] See Chapter 2.C.

[11] Vincent, (2021).

[12] United Nations Security Council, (2021), para. 63.

n't Uber Charged in a Fatal Self-Driving Car Crash?" is the title of another piece that highlights how those who developed an autonomous vehicle skipped the blame when it misclassified and killed a pedestrian.[13] Some other experts have already pointed out how robotic interrogators might cause ill-treatments —again, with no clear candidate to account for them. Remarkably, these challenges are the object of policy debates within the Council of Europe (CoE), whose European Committee on Crime Problems (CDPC) entrusted a working group with drafting an international instrument on AI and criminal law.[14]

In the face of these developments, the challenge lies in balancing the benefits of robots and AI while ensuring criminal accountability when things go wrong. The core remit of this thesis is to investigate the implications that robots have for the practices that determine that a person deserves criminal blame.

### B. Research questions

In light of such an overarching concern, the project reflects on three narrower research questions.

The first question is: to what extent is the attribution of criminal liability for robot's harm a matter of human rights? Some voices already defend giving preference to the benefits. In their opinion, either the latter outweigh the need to keep the bite of the criminal apparatus,[15] or it is more convenient to replace criminal law's blaming with some form of civil law's compensation.[16] Those voices invite reflecting on why and to what extent human rights standards demand keeping criminal liability.

The second question is: to what extent would the engagement of robots in a wrongdoing blur the assessment that a person is criminally liable? This question involves three sub-questions: (1) why would robots block the attribution of criminal liability in the view of scholars and organisations? (2) To what extent is it possible to refine the idea of criminal liability underpinning scholars and organisations' perspectives to better model the attribution of criminal liability for robot's harm? (3) Is it possible to deem a person criminally liable for the robot's injury without unduly restricting her rights nor frustrating accountability?

The third question is: what are the persisting reasons for concern in attributing criminal liability for the robot's harm? This latter question aims at tapping on previous reflections to identify those contexts where robots are likely to challenge accountability practices.

[13] Marshall, (2020).

[14] Gless, (2020), 15.

[15] See the accounts in Santoni de Sio and Mecacci, (2021), section 3.2.2.

[16] See Waxman, (2013), 17.

## C. Argument overview

The thesis starts with the simple yet not fully explored first question in Chapter 2: to what extent is the attribution of criminal liability a matter of human rights? The Chapter contests early proposals, focusing on the right to a remedy.[17] It then suggests zooming in a subset of human rights-based obligations, mainly developed within the ECtHR. Coined as duties of redress,[18] they justify an *obligation* to pursue criminal accountability whenever robots and AI go awry.

Chapter 3 asks: why would robots block the attribution of criminal liability in the view of some scholars and organisations? First, the Chapter models their vision of criminal liability and gaps in attribution. Second, it explains the arguments underpinning their views. Called "attribution gap theories," they reflect the prevailing position that either the device's unpredictability or autonomy, or the complex imbroglio of humans and machines behind it, would impede singling out someone as worthy of blame.

Chapter 4 suggests an alternative model for grasping those problems of attribution. The question here is: to what extent is it possible to refine the idea of criminal liability underpinning scholars and organisations' views to model better the attribution of criminal liability for robot's harm? The first step in outlining that idea is to show that it is also conceivable to deem an unaware subject liable for failing to realise the risks she was creating through a robot. The second step defends asking not whether someone is accountable, or not, but whether blaming that person meets the human rights prerogatives of both defendants and victims. It names it the "technologically blurred attribution" model to differentiate it from prevailing "attribution gap theories."

In light of that refined model, Chapter 5 asks: is it possible to deem a person criminally liable for the robot's harm without unduly restricting her rights nor frustrating accountability? The Chapter sets the model in action. It thus shows that, even if robots are unpredictable, complex or autonomous, it is possible to deem a person behind accountable without unduly restricting her human rights prerogatives nor rendering obligations of redress pointless. In this vein, it proposes a solution for the puzzles of attribution gap theories.

Chapter 6 continues with the following question: what are the persisting reasons of concern in attributing criminal liability for robot's harms? This Chapter uses the model of technologically blurred attribution to pin down two cases. The Chapter first argues that ML-based decision-assistance technologies would disrupt the manner users make decisions. The second case fo-

[17] See, e.g., Chengeta, (2020), 4-11.

[18] Mavronicola, (2017), 1027.

cuses on how autonomous robots might overstretch expectations of care on those who develop them.

The conclusion, in Chapter 7, briefly depicts the thesis's takeaways. It then draws some implications for the CoE's suggestion of an international framework.

### D. Methods and source materials

This section addresses the methods and sources chosen to solve the research questions posed above. As its heading indicates, this thesis is a disentanglement. It thus presupposes a perplexing or troublesome situation to untangle. How to pin it down? How to determine the entanglement that robots pose to criminal attribution when most problems are yet to exist? The thesis proposes a systematic review of the emerging literature on the topic of criminal attribution and robots.

The thesis undertakes such a review in three steps. First, an initial scoping that involves searching in six databases[19] with a predefined query.[20] The second step reviews the works to identify those focusing on deeming humans accountable for robots' wrongdoings. The thesis's interest in human rights and domestic criminal laws also justifies discarding literature that concentrates on IHL and international crimes. The last step is a manual review of the works referenced in the ones initially scoped. It identifies research in other languages and reports from organisations, particularly HRW and the CoE. The result was twenty-one scholarly documents and four reports drafted under the aegis of the two organisations above.

| | *Google Scholar*[21] | *Scopus* | *Oria*[22] | *Web of Science* | *PhilPapers* | *SSRN* |
|---|---|---|---|---|---|---|
| Scholarship | 50 | 47 | 50 | 34 | 4 | 1 |

*Figure 1. Results of systematic after the second step*

Now, the disentanglement is of the attribution of criminal liability. That implies a salient legal dimension. Given that criminal law is primarily a matter of domestic legal systems, the prob-

[19] Google Scholar, Oria, PhilPapers, Scopus, SSRN, and Web of Science. Reviewed between February and March 2021.

[20] ("Artificial Intelligence" OR "AI" OR "Machine Learning" OR robot* OR *bot) AND ("liability" OR "accountability" OR "responsibility" OR "attribution") AND ("criminal" OR "crime").

[21] Only the first 50 results were selected.

[22] Only the first 50 results were selected.

lem is how to address the question in a manner that is sufficiently general to cover different regimes and, at the same time, specific enough to have something meaningful to say.

As a solution, the thesis proposes focusing on international human rights standards.[23] They include obligations that are both specific and universal enough to inform different legal systems. But, how to unearth those standards if conventions rarely require mobilising criminal laws, and even when they do, they do not determine how states should allocate responsibility?[24] Treaty-bodies' and international courts' rulings offer a promising alternative. The thesis proposes Doctrinal Legal Research (DLR) to make sense of those decisions. The approach helps to synthesise the principles underpinning the rulings[25] and has the predictive power to grasp their relevance for tomorrow's technologies.[26]

The first step involves selecting a universe of cases. Two exclusions were necessary here. The first entailed excluding the realms of international humanitarian law (IHL) and international criminal law (ICL). In light of the study's interest in international human rights law, it leaves them aside and focuses on offences occurring in the ordinary functioning of modern societies. The implication is that it goes beyond gross, systemic and widespread violations, common in wars and situations of political unrest.

Even within that scope, there are right-threatening behaviours that seem foreign to the employment of robots or human rights standards. That leads to the second exclusion: cases like domestic slavery and human trafficking. As shown below, a review of offences likely to involve robots underpins the thesis, and none of the reviewed materials pointed to those behaviours. However, given the thesis's interest in the problem of attribution, which is common to all criminalised offences, its observations are in principle applicable within those contexts. The reader is only warned that the nuances of these cases are left aside.

The European Court of Human Rights (ECtHR) jurisprudence and that of the Human Rights Committee, the Inter-American[27] and the African[28] systems compose the universe of cases studied. The thesis focuses on two categories of decisions. First, those involving an obligation to regulate the attribution of criminal liability. As Chapter 2 argues, the allocation of criminal

[23] Technical standards and national rules also feature through the paper. However, they are not the direct source of the analysis.

[24] For some exceptions, see CPPCG art. VI; CAT art. 4; GCIV, art. 147.

[25] Hutchinson, (2018), 9.

[26] The Pearce Committee defines DLR as *research which provides a systematic exposition of the rules governing a particular legal category, analyses the relationship between rules, explains areas of difficulty and, perhaps, predicts future developments* (cited in Hutchinson and Duncan, (2012), 101.

[27] It includes the ICtHR and the ICTHR.

[28] It includes the ACtHPR and the ACHPR.

liability mainly depends on how the state shapes its regulatory framework, whereas obligations to prevent, investigate or punish are secondary. The second category is those decisions defining the requirements to allocate criminal liability. The thesis seeks to untangle the problem of robots impeding to attribute harm back to a user or developer "behind" them. Hence, the interest in those requirements is mainly covered within the right to a presumption of innocence.

Because of that focus, the ECtHR's jurisprudence bears prominence. The latter has developed a more nuanced set of obligations and has faced a more varied scope of cases. The HRCtte, the ICtHR and ACtHPR are mainly referenced to put the latter in the context of a trend. Overall, this study intends not to be overly descriptive and exhaustive but to observe how human rights standards could inform the deployment of robotics and AI.[29]

The second step of the DLR is extracting principles that justify and bring order to judicial propositions.[30] However, it involves more than fetching any explanation. The principle's acceptance hinges on moral conceptions, the latter supported by interdisciplinary social science research.[31] In this sense, the thesis interprets the jurisprudence considering research about the function of criminal liability and the minimal conditions to reasonably attribute liability to an inadvertent subject.

Apart from the interdisciplinary dimension of DLR, the piece integrates knowledge from other disciplines to substantiate claims —that is, for heuristic purposes—[32] or to identify problems that may require legal responses —known as auxiliary uses.[33] Both the third research question and the second sub-question to the second query —to what extent is it possible to refine the idea of criminal liability underpinning the view of scholars and organisations? — use the model of legal disruption[34] with heuristic purposes. Indeed, it justifies zooming into the individual "behind" the robot and improving the definition of cases where technologies block attribution. Contesting the dominant positions also involved the heuristic use of notions like "cognitive uncontainability" to better grasp the unpredictability that robots pose.[35] In

[29] For descriptive studies, see: Seibert-Fohr, (2009); Kamber, (2017).

[30] Bhat, (2020), 159-61.

[31] Dworkin, (1973), 102(2); Van Hoecke, (2011), 1-18, 4-7.

[32] See van Klink and Taekema, (2011), 3-6. They distinguish five models of interdisciplinary research, with increasing levels of integration: heuristic, auxiliary, comparative, perspectivist and integrated.

[33] Ibid.

[34] Liu et al., (2020).

[35] See Chapter 5

turn, the thesis uses the idea of “face validity” with the auxiliary purpose to understand how AI’s non-intuitiveness blurs the assessment that a user is criminally liable.[36]

Finally, the thesis also sought to scope those emerging and prospective cases of robots involved in crime. Here, it relied on the reviews already conducted by other authors, chiefly those of Hayward and Maas,[37] on the one hand, and King et al.,[38] on the other. The thesis focuses on what Hayward and Maas have identified as “crimes by AI.”[39] It refers to those instances where the device performs the offence with no involvement of a user. The reason is that those are the kind of cases that supposedly present a gap in criminal attribution.

### E. Status of the subject matter and contributions

The first question —to what extent is the attribution of criminal liability for robot’s harm a matter of human rights? — has been answered before in light of the right to a remedy.[40] Ensuring criminal accountability of individuals —be it in domestic regimes or at the international community level— is perceived as essential to victims of violations of IHL and human rights standards.[41] Insofar as robots impede that accountability, their deployment “should be considered unethical and unlawful.”[42]

Those concerns have focused on lethal autonomous weapons (LAWS). Only recently, some scholars have started to identify the problem as one applicable, not only to weapons but to a broader array of intelligent and connected artefacts.[43] It has also surfaced as a policy issue within the CoE[44] and the Australian Human Rights Commission.[45] These are the first two institutions to point to a human rights-based obligation to ensure clear legal rules regarding liability for the use of AI.

Despite the scholarship and the emerging policy discussions, there is one crucial blind spot: the obligation to keep criminal accountability, not only in cases of political unrest, armed con-

[36] See Chapter 6.

[37] Hayward and Maas, (2021).

[38] King et al., (2020).

[39] Hayward and Maas, (2021), 217-18.

[40] See Chapter 2.

[41] See, e.g., Chengeta, (2020), 4-11.

[42] UNGA, (2013), para. 80.

[43] Rodrigues, (2020), Liability for damage.

[44] *Member States must ensure that effective remedies are available under respective national jurisdictions, including for civil and criminal responsibility, and that accessible redress mechanisms are put in place for individuals whose rights are negatively impacted by the development or use of AI applications* (CAHAI), (2020), 39.); See also Council of Europe, (2021), 21.

[45] Australian Human Rights Commission, (2021), 79.

flict or state-sponsored crime but also for deaths and injuries occurring in times of peace. The criminal law literature has mainly focused on LAWS in contexts of war and political unrest. Simultaneously, those going beyond these technologies fail to zoom into the peculiarities of criminal liability. To what extent are states obliged to keep the bite of their criminal apparatuses when an autonomous car crashes on a pedestrian or a robotic surgeon injures a patient remains unaddressed. The problem becomes serious if one considers that some voices are suggesting to give up with punishment and accountability in dealing with new technologies. [46]

Filling that gap is the piece's first contribution. It discusses to what extent states are obliged to keep criminal attribution for the harms that robots and AI —and not only LAWS— might cause in functioning societies. In so doing, it develops and gives substance to developing policies that aim at ensuring individual accountability for the robot's harms. As the piece shows, the particularities of criminal law call for a specific justification; one that might differ from that of other regimes, like torts or product liability.

The second and third question point directly to the problem of accountability. Whereas scholars and policymakers have raised issues of civil, torts and product liability quickly, the problem of *criminal* accountability has largely lagged behind.[47] In the former, the problem is identifying the cheapest cost avoider or insurance taker, regardless of their involvement.[48] In this sense, there seems to be consensus around the convenience of combining faultless and joint liability with changes in the burden of proof.[49]

However, those solutions do not play well for criminal liability, which still requires identifying a blameworthy cause. In this latter field, the dominant position among the few scholars is that some autonomous robots —especially those exhibiting AI or self-learning capabilities— will pose an intractable liability gap. Such a gap means that there will be cases where the use of a robot will entail that no one can be fairly deemed responsible.[50]

Can one be confident about a criminal attribution gap if an autonomous vehicle misclassifies and knocks a pedestrian down, as it happened in Texas less than three years ago?[51] If an AI-based interrogator causes great suffering on the interrogatee, could one say that no one would

[46] Arguing that such an idea does not reflect anything more than "an a priori principle that there must always be a human to hold accountable," see:  Waxman, (2013), 17.

[47] Within the EU, see e.g.: Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics, (2017); European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence, (2020).

[48] See also European Parliament Committee on Legal Affairs, (2017).

[49] See, e.g., Bertolini, (2020).

[50] Lagioia and Sartor, (2020), 434.

[51] See Marshall, (2018); Uber's Self-Driving Operator Charged over Fatal Crash, (2020).

answer for the offence of ill-treatments?[52] As soon as one scratches the surface, the problem becomes a thorny one. This piece can be seen as an original contribution to an ongoing debate. It contests the prevailing position that robot's unpredictability, complexity and autonomy block criminal attribution. It also proposes a model to attribute liability and, at the same time, capture those instances of blurred attribution. Finally, it points to two areas where robots pose a gap that scholars seem to be nonetheless ignoring.

Apart from filling gaps in the literature and presenting an alternative opinion, the thesis contributes to an emerging policy discussion. At a domestic level, countries like France are introducing laws with penal consequences for the use of autonomous vehicles.[53] Additionally, the Singapore Academy of Law Reform Committee issued its "Report on Criminal Liability, Robotics and AI Systems" earlier this year.[54] The report highlights important issues around the problem of attributing criminal liability for robots and goes on to suggest alternative methods to the use of criminal law.[55]At the international level, the CDPC held a thematic session on AI and criminal liability back in 2018.[56] Following such a session, a working group carried out a "feasibility study" on a future international instrument on AI and criminal law.[57] After the study, the working group began last year to draft such an instrument.[58] Overall, the thesis discusses topics central to those policy initiatives and, thus, adds its own appraisal to an emerging discussion.

### F. Caveats

This thesis has nonetheless some limitations worth stressing from the outset. The first is the thesis's focus on criminal *attribution*. The latter determines that wrongdoing is someone's work and is necessary to convict someone. Nevertheless, it is not a sufficient one. The mobilisation of the criminal apparatus involves many more aspects, like evidence, procedural rights and defences. Those remain nonetheless out of the piece's scope. The problem of attribution itself is a complex one and deserves specific attention. Furthermore, issues like evidence and AI are so complicated —and sometimes contextually dependent— that they call for their own consideration.

[52] See McAllister, (2017).

[53] Law no. 2019-486 of 22 May 2019 on the Growth and Transformation of Companies, Art. 125 modifying ordinance No. 2016-1057 on the Testing of Vehicles with Delegation of Driving in Public Roads.

[54] Law Reform Committee, (2021), 13-17.

[55] Ibid., 44-45.

[56] CPDC, (2018).

[57] Gless, (2020).

[58] Ibid., 15.

The second caveat has to do with the international focus. The thesis does not zoom into the specificities of any national law, nor it is a comparative law assessment. It instead taps on international human rights standards to disentangle the problem of attribution gaps. In this vein, domestic laws and technical standards are mainly the focus of examples or arguments that ground the analysis; although, they are not the primary target. Future research could use the model developed here to zoom into those domestic norms and industry guidelines.

The third caveat concerns the focus on criminal law within functioning societies in times of peace. Its main interest is building on human rights standards to solve the problem of criminal accountability. The thesis does not ignore the challenge of LAWS to IHL and liability for state-sponsored crimes. However, it leaves them aside to focus on the problems that increasingly ubiquitous robots pose outside warfare contexts and social unrest. Would CAVs, robotic interrogators or autonomous surgeons blur criminal accountability in the more familiar setting of an ordinary society in times of peace? The thesis goal is to offer that concern the separated attention it deserves.

## G. Definitions

### *i. Criminal liability?*

The ECtHR’s definition of criminal law is helpful to pinpoint what is specific to the regime. In this vein, the thesis uses the criteria set out in *Engel and Others*.[59]

It considers that a judgement pertains to criminal liability insofar as:

(i) Is so designed in the domestic law at stake;

(ii) the behaviour being of such a nature that it demands a deterrent response, and[60]

(iii) the consequence either follows a condemnation in a criminal procedure or demands being treated as a penalty in light of its punitive aim, its classification under domestic law, its execution and severity. [61]

### *ii. Liability, attribution, responsibility and accountability*

*Responsibility* and *accountability* are connected notions. To be responsible is to be open to being held or called to account for an alleged failure to act as one should have acted.[62] Re-

[59] *Engel and Others v. the Netherlands*, no. 5100/71; 5101/71; 5102/71; 5354/72; 5370/728, § 82, ECHR, 1976, Series A no. 22.

[60] *Jussila v. Finland* [GC], no. 73053/01, § 38, ECHR 2006-XIV.

[61] *GIEM SRL and Others v. Italy* [GC], nos. 1828/06 and 2 others, § 211-219, ECHR, 2018.

sponsibility, thus, points to a certain status where someone is susceptible to being asked to give reasons for failing to meet certain expectations attached to a role. Accountability, in turn, points to the situation of being called to answer for such a failure. Being accountable thus presupposes a standard of conduct. Someone is accountable *for* failing to conform to certain norms that dictate how one is to behave.[63]

Accountability has two aspects: *answerability* and *liability*. To be liable means that an adverse event is imputed to the behaviour of a person. It points to the satisfaction of the conditions for being deemed accountable *for* failing to meet some expectations. In turn, being answerable means being accountable *to* someone else. When a defendant, for instance, appears to a court and denies that the court has jurisdiction to adjudicate on her matter, she does not deny her liability but her conditions of responsible vis-à-vis the specific institution.

In this sense, while liability presupposes answerability (an agent A is liable to blame, or conviction, for φ only if A is answerable), answerability does not entail liability (A can admit that is answerable but avert blame or conviction by offering a persuasive justification or excuse for an action φ).[64]

To facilitate the discussion, the term "attribution" is kept to single out those circumstances where one can say that an agent A owns some φ-doing. It points to the connection between an agent and a behaviour that is hers.[65] What seems to interest the prevailing position is that robots frustrate such a judgement, whereas issues of answerability and responsibility are dealt as separate problems.[66] Hence, the focus of this piece is on *attribution*. The latter might be a necessary condition to say that someone is liable but is not sufficient. Indeed, liability might add further requirements depending on the legal system.[67] Dealing with those requirements would have entailed a broader comparative analysis that nonetheless would have missed the problem that prevailing positions are highlighting.

[62] Watson, (1996). It is worth stressing that such responsibility also cover good deeds. Moreover, it can be prospective or retrospective. The first category points to a position where someone is open to be called to account for complying with a set of obligations attached to the role. In turn, the second point to the situation where someone is open to account for the way one discharged an obligation. The difference here is that, whereas retrospective obligation is backward-looking in that it seizes past conducts considering the set of norms, prospective takes into account future conducts (ibid.).

[63] Duff, (2018), 776-77.

[64] Ibid., 777.

[65] Cf. Hart's notion of capacity-responsibility in Hart, (2008), 227-30.

[66] Distinguishing different types of responsibility, and highlighting how most accounts focus only on backward-looking attribution of some harm, see: Santoni de Sio and Mecacci, (2021).

[67] Hart, (2008), 215-30.

### *iii. Robots and AI*

There is some consensus around the idea that robots are physical objects that take the world in, process what they sense, and in turn act upon it. In other words, the sense-think-act paradigm. Robots are thus cyber-physical systems (CPS). That means that they combine a sort of physical embodiment with computing power and actuators to modify the external world. [68]

In a nutshell, a robot is a moving machine controlled by a “computer.” That “computer” could certainly be a magnetic drum imparting step-by-step commands to a mechanical arm, as the early Unimate.[69] But, as robots go on to pullulate in more unstructured environments, more powerful computers become indispensable.

Here is where artificial intelligence —AI— joins the stage. As a scientific discipline, it is concerned with making machines *intelligent*, where “intelligence” is that quality that enables an entity to function appropriately and with foresight in its environment. More precisely, AI is the scientific study of what problems can be solved, what tasks can be accomplished, and what features of the world can be understood, and then to set forth procedures —that is, algorithms— to show how this can be done efficiently.[70]

Seen from the perspective of the result, AI is algorithms and data. Simply put, it is a procedure to solve a problem and the information needed to do it. What singles out AI from other algorithms is its ability to select an action, among a scope of alternatives, that would maximise a performance measure. Then, there is nothing like human intelligence in AI or robotics. Combined, they yield a set of data and problem-solving procedures that steer a machine to be efficient at solving a problem.[71]

The “how question” inevitably leads to machine learning (ML). It literally means that the machine teaches itself the “correct,” or rather “useful” rules, it needs to perform effectively. It does not mean acquiring new concepts, as a child would. Instead, it means that, when supplied with a large amount of data,[72] the machine improves itself in finding a function to map the features of the input data to the desired outputs. Instead of a child acquiring new knowledge, ML “learns” as an extremely successful statistician would: it takes in information, and, iteration after iteration, optimises the way it detects patterns in that information.

[68] Calo, (2015), 529-32.

[69] See Gasparetto and Scalera, (2018).

[70] Maas, (2021), 34-39.

[71] Ibid.

[72] McQuillan, (2018), 2-4.

Robots and AI are good partners, but they are not essentially connected. It is possible to have a robot with a “computer” that does not involve any AI technique. At the same time, many AI applications do not embody a machine. Instead, they operate “through” disembodied agents. AI does not need a body, nor the machine needs an AI-mind. For the sake of the present exposition, the words “robot” and “AI” are used as a synonym that covers these overall spectra of technologies. If necessary, the piece spells the features that are relevant for each point.

Key for the piece, however, is ML’s ability to find patterns beyond human intelligence. The claim has been at the roots of AI’s inscrutability, yet it chiefly applies to “neural networks,” also known as deep learning.  The deep learning algorithm itself consists of different layers of nodes or “neurons” emulating a human brain. Each one is connected to all the nodes in the subsequent layer, each node-to-node connection representing different weights.[73] From there not only AI’s ability to solve complex problems but also the difficulties for a human to grasp the machine’s reasoning.

[73] Ibid., 3.

# 2 A human rights obligation regarding allocation of criminal liability?

"An understanding of how to protect human rights in the digital context is significantly underdeveloped."[74] The phrase is Eileen Donahoe's, former US Ambassador to the Human Rights Council and director of global affairs at Human Rights Watch. And it applies to the problem of criminal liability sketched in the introduction to this piece. There is a growing uneasiness around the idea that emerging technologies will introduce gaps in attribution.[75] And yet, as one scratches the surface, there are no detailed accounts on why it is a human rights problem.

That issue is not out of practical importance. If such an attribution gap does not infringe any human rights standards, states are in principle free to introduce robots without failing to meet their obligations.[76] Should the opposite be the case, however, apportioning liability would not be a mere option. States would be obliged to patch those legal gaps as emerging technology enters the society—or restrict their introduction when it proves impossible.[77] Institutions and practices of responsibility are critical, at least within contemporary constitutional democratic orders.[78] But do such practices also involve human rights standards? If so, how could states deal with the challenges that robotics and AI are introducing?

Early reverberations stress the relevance of existing entitlements to a remedy. Section A engages with those approaches. It suggests shifting the attention to a subset of duties, mainly developed under the aegis of the ECtHR. Those duties are of interest because they oblige states to mobilise the criminal law against specific human rights violations. Section B provides a detailed account of them. First, it delineates the kind of transgressions they tend to tackle. Then, it moves to define the type of obligations that they impose. Lastly, Section C canvases their rationale and relevance for the problem of attributing machine harm.

## A. From remedies to duties of redress: why attribution gaps contravene human rights?

This Section engages with those approaches that have pointed out to the right to a remedy. It presents them in Sub-section i. and introduces the objections in Sub-section ii. As the Section

[74] Donahoe, (2016).

[75] See, e.g., Gless, (2020).

[76] See Law Reform Committee, (2021), 45-46.

[77] Cf. Hildebrandt, (2008). In her view, *if we cannot attribute criminal liability for wrongful actions because the responsibility is diffused beyond measure, we should think twice before introducing the technological infrastructure that enables such unaccountable consequences* (178).

[78] Council of Europe, 8.

argues, a different approach is necessary to account for the specificities that criminal law brings in.

### *i. Initial approaches*

Whenever robots' nature renders responsibility for its consequences impossible, "its use should be considered unethical and unlawful."[79] Expressions like that featured in early accounts of the human rights challenges that robotics might bring in. The first attempt to pinpoint those issues came with a report that Christof Heyns submitted to the UN Human Rights Council.[80] Focusing on autonomous weapons, the then special rapporteur on extrajudicial killing grounded his argument on the importance of attribution for the overall effectiveness of human rights. Instead of identifying a specific prerogative, Heyns looked at the importance of attribution for safeguarding human rights' bite in the digital era.[81]

Later approaches focused instead on the right to a remedy.[82] The assumption is that the right to a remedy requires states to ensure individual accountability, a kind of accountability that robots will frustrate. The idea first appeared regarding autonomous weapons[83] but further expanded to other types of robots. In their report for the Council of Europe, for instance, Rinie van Est and Joost Gerritsen pointed to such a right in the context of autonomous vehicles.[84] Again, the Australian Human Rights Commission referred to the right to an effective remedy as the source of a human rights obligation to patch the attribution gaps that AI will introduce.[85]

Focusing on criminal issues around LAWSs, one scholar argues that the right to a remedy shelters an obligation to investigate human rights violations and bring perpetrators to justice through prosecution.[86] In his opinion, such a right has three components: victim's access to justice, reparation[87] and the right to access information and to know the truth concerning the infringement of their rights.[88] Combined, the three components oblige states to prosecute of-

[79] UNGA, (2013). para. 80.

[80] Ibid.

[81] Ibid. See, for instance, para. 75.

[82] HRW and IHRC, (2014).

[83] Ibid.

[84] Rathenau Instituut, (2017), 33-37.

[85] Australian Human Rights Commission, (2021), 73-80.

[86] Chengeta, (2020), 7. He also adds that individual accountability is a matter of customary international law. However, he seems to refer here to crimes against peace (aggression), war crimes, genocide and crimes against humanity (see ibid., 16; cf. Swart and Cassese, (2009), 82.).

[87] Cf. Seibert-Fohr, (2009), 40.

[88] UNGA, (2005), para. 24.

fenders and ensure that non-state actors provide reparations upon their conviction.[89] Other scholars have followed a similar approach. [90]

In her summary of current debates, however, Rowena Rodrigues adds the right to life along the right to a remedy. She believes that failures of attribution directly frustrate the rights of victims to obtain a remedy. At the same time, however, such failures undermine the protective bite of other rights—like the right to life. [91]

Andrea Bertolini went one step forward, however. Focusing on liability for defective products, he looked at the technology-chilling effect of uncertain liability frameworks.[92] His point is that such an unreliable framework contravenes certain human rights obligations. Specifically, it infringes states' duty to promote research and developments of devices for persons with disabilities, as included in the CRPD.[93]

In sum, attribution failures engage both the right to a remedy and the general human rights framework. On these grounds, it is possible to distinguish between a *direct* and an *indirect* approach. The first assumes that the right to an effective remedy entails an obligation to attribute behaviour. A commitment that robots would frustrate should they pose an attribution gap. The second approach turns the focus to the possible consequences that such a gap might pose. Instead of pointing to the infraction of an obligation, the focus here is on the effects that failures of attribution might bring in.

In this sense, the second approach is empirical. The question is not whether the state failed to do something it ought to, but what would be the upshot of such failures. When Andrea Bertolini asks if product liability failures might undermine the rights of persons with disabilities, he pinpoints the technology-chilling effects of such defects. In this sense, he takes the stakes of the debate from the normative to the empirical field: gaps in attribution lead developers in the external world to refrain from developing robots. That statement is empirically falsifiable since one could always offer further evidence that that is not the case. If accepted, however, the conclusion is that it violates some rights, like the stimulation of research and technology for persons with disabilities.

[89] Chengeta, (2020), 4-11.

[90] See, e.g., Koops et al., (2013).

[91] Rodrigues, (2020).

[92] Bertolini, (2015), 126-30.

[93] CRPD, Article 4 (g).

### *ii. A right to a remedy?*

The accounts above succeed in pinpointing the human rights salience of attribution gaps. Furthermore, they make clear that technologies are never created in a legal vacuum. They come into an environment where international human rights standards have something to say. And trying to make that discourse plain is a commendable effort. At the same time, however, they fail to justify *why* failures to *apportion* criminal liability pose a human rights issue.

To begin with, it is not clear from the right to a remedy that criminal attribution is pertinent. At least, that is not the case beyond the scope of serious abuses. Consider HRW's report on LAWS and law enforcement.[94] In making its claim that attribution problems engage the right to a remedy, HRW quotes UN's Basic Principles and Guidelines on the said right.[95] However, those guidelines deal with gross violations, amounting to international crimes.[96] It thus begs the question of other breaches that, even if severe, do not entail crimes against peace, war crimes, genocide or crimes against humanity. What to do with more mundane cases of robots interrogating passengers in borders? Are the standards quoted in HRW's report also applicable when an autonomous vehicle knocks down a pedestrian?

In any case, the kind of attribution gaps that robots might pose is mainly a problem of how laws are shaped. Vacuums in responsibility do not arise from failures to investigate or to bring potential perpetrators to justice. Instead, they have their roots in law's failure to cushion the kind of problems novel technologies will introduce. A victim might know the truth about the wrongdoing, and she might see developers or users in the bench. And yet, they might not be criminally responsible. Hence, the approach ill-suited to address the type of attribution problems AI and robotics will pose. Whoever is behind robot harm will skip blame, not because of lack of willingness to investigate her, but because she cannot be responsible for such injury. And the reason why she cannot be is that laws are ill-suited to make her liable.

Lack of verification is the problem when it comes to the indirect approach. On the one hand, it is unclear whether uncertainty will generate the kind of technological-chilling effects that Andrea Bertolini forecasted in his paper.[97] Conversely, the indirect account focuses on product liability.[98] In this sense, it does not explain why a failure to apportion *criminal* liability

[94] HRW and IHRC, (2014).

[95] UNGA, (2005).

[96] Ibid. The Resolution's Preamble stresses that the "Basic Principles and Guidelines contained herein are directed at gross violations of international human rights law and serious violations of international humanitarian law which, by their very grave nature, constitute an affront to human dignity." (Ibid., Preamble).

[97] Bertolini, (2015).

[98] Ibid.

will undermine the human rights regime. And in light of the current evidence at stake, it seems too weak as an argument.

Overall, these theories fail to differentiate criminal liability from other kinds of responsibility. As Chapter 1 clarified, "responsibility" has different meanings depending on the context. Legal systems typically comprise several layers of responsibility. Once something goes wrong, the victim might introduce a civil suit to get compensation or submit a claim leading to a disciplinary sanction. Criminal law is one of those layers, and, in contrast to some other alternatives like product liability, it does not aim at distributing economic burdens. Criminal liability is concerned with allocating blame for undesirable behaviours. As such, it seeks to single out an agent and communicate blameworthiness for whatever she did.[99] Thus, explaining why human rights demand a *criminal* response to robot harm requires a more specific approach.

"Duties of redress" offer a promising option. With such a term, Natasa Mavronicola[100] refers to a set of positive obligations demanding, not a remedy, but the pursuit of criminal redress. Often arising from the rights to life and privacy, as from the prohibition of ill-treatment, they offer several advantages to frame the gaps robots might bring in.[101]

Certainly, there are overlaps between duties of redress and the right to a remedy. Those overlaps, for instance, have led the ECtHR to acknowledge that an examination of a matter under those duties makes it unnecessary to further address them in light of the right to a remedy.[102] However, duties of redress have the advantage of entailing an obligation to modify the legal framework. It is not by chance that, in referring to them, George P. Fletcher concluded that the ECtHR had assumed "the remarkable burden of supervising and rewriting the criminal codes of all the member states."[103] Since such a "rewriting and supervising criminal codes" is what an attribution gap demands, they offer an unparalleled vantage point for analysing the future robots will bring in.

[99] See Sub-section iii. And Chapter 4.

[100] Mavronicola, (2017), 1027.

[101] Ibid. Although not expressly provided under the ACHPR, both the ACtHPR and the AComHPR read in the substantive provisions (see, e.g., AComHPR (Communication) *The Social and Economic Rights Action Center and the Center for Economic and Social Rights v. Nigeria*, no. 155/96, 2001, § 68). However, often linked with securing further remedial measures for the victims. The HRCtte construes the obligation based on article 2 (3) (the right to a remedy) taken in conjunction with one of the substantive provisions of the ICCPR (See HRCtte, no. 972/01, *Kazantzis v. Cyprus*, CCPR/C/78/D/972/2001, 2003, § 6.6.). In turn, the ICtHR builds upon Articles 1(1) (obligation to respect rights) and 2 (domestic legal effects) of the ACHR (IACtHR, no. 7920, *Velásquez Rodríguez v. Honduras* (Judgment), 1988, § 162-163). See also Kamber, (2017), 122-24, 70, 89-90.

[102] See, e.g., *X & Y v. The Netherlands*, no. 8978/80, § 36, ECHR, A91.

[103] Fletcher, (2005), 553.

### B. Defining duties of redress

This Section delineates the “duties of redress” sketched above. It starts with an account of its triggers to determine *when* robot harm will yield an obligation to adapt the legal framework. It then turns to the core of the obligation and asks *what* states must do under those duties. To what extent are they under an obligation to “rewrite” their codes, as the Cardozo Professor of Jurisprudence pointed out? Lastly, the section explains the motivation behind those duties. The point is relevant to answer *why* criminal law, but not compensations or any other remedy, is the preferred alternative to deal with robot’s harms.

#### *i. When do duties of redress enter into play?*

Regarding duties of redress, it is worth it to start by asking what kind of events give rise to such obligations. When are states obliged to deploy their criminal apparatus such that a failure to do so might engage their international responsibility?

It is possible to distinguish two kinds of triggers. On the one hand, there are “gravity triggers.” These triggers follow the severity of certain situations. More severe events will demand a criminal response, whereas other alternatives will suffice for less severe cases. On the other, it is possible to identify “equivalence triggers.” Whenever states have opted for criminal law to deal with certain situations, gaps based on the victim’s status are unacceptable. Hence, it is not the gravity that grounds the state’s responsibility, but the sort of inequality arising from leaving some victims outside the range of criminal law.

Gravity triggers

Gravity triggers are the first sort of events giving rise to duties of redress. What kind of events do they encompass? On one end of the spectrum, there are life deprivations and life-threatening injuries. Killings and serious injuries are so severe that they oblige the state to mobilise its criminal law apparatus to protect and redress the victims.[104]

As an example, in *Norbert Zongo and Others* —concerning the assassination of a prominent Burkina Faso journalist— the ACtHPR found a violation of Article 7 (access to justice) because the state “did not acted with due diligence in seeking out, prosecuting and placing on trial” those responsible.[105]

[104] See, e.g., *Estamirov and others v. Russia*, no. 60272/00, § 85-87, ECHR, 2006; *Yasa v Turkey*, no. 22495/93, § 98-100, ECHR, 1999-VI; See *Velásquez Rodríguez* § 162-163; AComHPR (Communication), *Mouvement Burkinabé des Droits de l’Homme et des Peuples v Burkina Faso*, no. 204/97, 2001.

[105] ACtHPR, *Beneficiaries of the Late Norbert Zongo and others v. Burkina Faso,* no. 013/2011, 2014, § 156.

The circumstances giving rise to the duty, however, are not completely clear. There are significant differences, at least in what concerns the HRCtte and the ECtHR. For the former, duties of redress cover all manifestations of violence or incitement, including intentional and negligent death.[106] The Strasbourg Court follows a more restrictive approach. Under the latter, criminal responses to negligent killing are the exception. Those exceptional circumstances include cases where, beyond an "error of judgment or carelessness," the harm results from disregard for right-threatening risks. [107] Besides those exclusions, the general rule is that states can safeguard unintentional injuries with civil remedies.

Consider, for instance, the seminal case of *Öneryildiz*. On 28 April 1993, a methane explosion in the Ümraniye municipal rubbish tip engulfed ten slum dwellings, killing thirty-nine people.[108] The Court, sitting as Grand Chamber, considered that the state knew and yet neglected the risk of an explosion for several years. And it did so despite several recommendations and reports.[109] It is considering such a disregard for the risks that the Court deemed administrative remedies insufficient. The Court then zoomed into the criminal investigation, which in its view, failed to meet human rights standards. Indeed, it considered that an investigation focused on offences concerning careless performance of duties fell short of making life-endangering behaviour the object of blame.[110]

Contrast the case above with *Calvelli and Ciglio*. When examining a case of medical negligence leading to the decease of a new-born, the Grand Chamber decided that the state needed not to respond with the apparatus of criminal law.[111] In truth, cases of unintentional deaths leading to a duty to offer criminal redress are rare in the ECtHR jurisprudence. Apart from *Öneryildiz*, the Court has only considered road safety,[112] transportation of dangerous goods,[113] denials of healthcare,[114] and military activities[115] as sufficiently severe to demand a criminal response.

[106] HRCtte, *General Comment No. 31: The Nature of the General Legal Obligation Imposed on States Parties to the Covenant*, 80th session, 26 May 2004, CCPR/C/21/Rev.1/Add. 13, para. 8.

[107] *Öneryildiz v. Turkey*, no. 48939/99, § 93, ECHR, 2004-XII.

[108] Ibid, § 18.

[109] Ibid. § 102.

[110] Ibid. § 116. Cf. *The Social and Economic Rights Action Center and the Center for Economic and Social Rights v. Nigeria* (2001), regarding a destructive oil explosion.

[111] *Calvelli and Ciglio v. Italy*, no. 32967/96, § 51, ECHR, 2002-I.

[112] *Smiljanić v. Croatia*, no. 35983/14, ECHR, 2021.

[113] *Sinim v. Turkey*, no. 9441/10, ECHR, 2017.

[114] *Asiye Genç v. Turkey*, no. 24109/07, ECHR, 2015.

[115] *Oruk v. Turkey*, no. 33647/04, ECHR, 2014.

That the HRCtte follows a broader approach becomes plain *Novaković*.[116] Similar to *Calvelli and Ciglio*, it concerns allegations of death by medical malpractice. In contrast to the ECtHR's decision, the HRCtte held that the state was obliged to investigate and, if appropriate, prosecute those responsible.[117] Particularly, the Committee rejected the possibility that administrative disciplinary or other remedies could satisfy the state's obligation.[118] Similarly, the ICtHR's considered that the state failed to meet its obligations by not effectively investigating and extraditing the potential responsible in a case of medical malpractice.[119]

The distinction is important for robot's harm. As Chapter 4 argues, apportion gap cases involve situations where the human behind did not foresee a harmful outcome. Think of an autonomous vehicle that runs over a pedestrian without any human behind intending or being aware. Is it a mere error of judgment so that the state does not need to involve criminal responses? Or does it entail a case of gross negligence? It is impossible to draw general distinctions here. It is thus sufficient to notice that, if one follows the ECtHR's jurisprudence, cases involving robots might raise debates on what counts as gross negligence. As Chapter 6 shows, AI precisely disrupts determining whether an inadvertent agent was careless. Consequently, one can expect the assessment of whether a display of carelessness was "sufficiently gross" to be also difficult.

Below the cluster of life-endangering behaviours, states are still obliged to criminalise other conducts, even if they do not put lives at risk. These conducts often fall under prohibitions of ill-treatment and thus need to be of a certain harshness.[120] For the ECtHR, what counts is whether the act was "premeditated" and caused "either actual bodily injury or intense physical and mental suffering."[121] Moreover, the nature and context and duration of the treatment are also relevant.[122] Last but not least, it is an obligation sensitive to the vulnerability and gender of the victim.[123] Indeed, the Court also weighs the victim's sex and susceptibility, including his relationship with the ill-treatment author.[124]

[116] HRCtte., no. 1556/2007, *Marija and Dragana Novaković v. Serbia*, CCPR/C/100/D/1556/2007, 2010. C

[117] Ibid., § 7.3.

[118] Ibid., § 6.3.

[119] ICtHR, no. 12406, *Albán-Cornejo et al. v. Ecuador*m, 2007, § 109.

[120] When it comes to acts of torture, the CAT obliges states to mobilise their criminal apparatus which encompasses criminalisation, asserting jurisdiction and prosecution of such acts (Article 4). In turn Article 16(1) — referring to cruel, inhuman or degrading treatment or punishment— seems to leave open the question of whether criminal law should be involved. See: Kamber, (2017), 191.

[121] *Beganović v. Croatia*, no. 46423/06, § 65, ECHR, 2009; *Sabalić v. Croatia*, no. 50231/13, § 64, ECHR, 2021; *A. v. The United Kingdom*, no. 25599/94, § 22-24, ECHR, 1998-VI.

[122] *Beganović* § 64.

[123] "Victim" is a vague notion, with its exact meaning largely depending on the context. Cf. UNGA, (2005); de Casadevante Romani, (2012).

[124] Ibid.; *Volodina v. Russia*, no. 41261/17, § 73, ECHR, 2019. See also Heri, (2020).

The seminal case within this prong, *A. v. The United Kingdom*, dates from 1998. In *A*, the Court considered that United Kingdom's defences allowed parents to get away without punishment for beating their children.[125] That context is not different from what the HRCtte comparatively found in *Rajapakse*, on that occasion after an arrest. [126] Thus, it seems that particularly intense cases of distress or pain can give rise to an obligation to deploy criminal redress.

Think of a sex robot that deviates from its programming in a manner that a user cannot stop it from engaging in coitus.[127] That case is not different from *MC v. Bulgaria*, which involved rape without physical force.[128] There, the Court reasoned that both the judiciary and prosecutors interpreted the crime of rape so restrictively that it failed to offer criminal redress. Would it not be the same for a state that allows commercialising autonomous sex robots without unblocking the path for criminal liability? If the robot behaviour is so unpredictable that current laws do not allocate liability, the state arguably will fail to comply with its international obligations. One might also apply the same principle where domestic authorities fail to redress victims of violent suppression of peaceful demonstrations[129] or when using a robot interrogator leads to acts amounting to torture or ill-treatments.[130]

Lastly, some violations of privacy also give rise to an obligation to criminalise. *X & Y*, the Court's first case on duties of redress, is on unconsented sex with a person with a mental health condition —that the Court deemed an infringement upon the right to privacy. Within the ECtHR, the key to unlocking the criminal alternative is that "fundamental values and essential aspects" [131] of private life are at stake. These fundamental aspects are unclear and demand a debate falling outside the scope of this piece. However, it is worth noting that an interplay between the victim's vulnerability and the intensity of the harm justifies recurring to the criminal apparatus. Others, particularly the ICtHR, are less specific. The latter has deemed states responsible for falling to investigate offences arising from the interception of phone conversations.[132] However, it has not specified whether *any violation* of the right to privacy should lead to mobilising the criminal apparatus.[133]

[125] *A.* § 24.

[126] HRCtte, no. 1250/2004, *Rajapakse v. Sri Lanka*, CCPR/C/87/D/1250/2004, 2006, § 9.5. Regarding the African system, see Basch, (2008).

[127] Currently, those concerns mainly appear in tabloids (see Moran, (2019). Still, the scenario is not at all unlikely.

[128] *MC v Bulgaria*, no. 39272/98, § 166, ECHR, 2003-XII (extracts).

[129] See AComHPR (Communication) *Kevin Mgwanga Gunme and others., v. Cameroon*, no. 266/03, 2009, § 112.

[130] AComHPR (Communication) *Egyptian Initiative for Personal Rights & Interights v. Egypt*, no. 323/06, 2011, § 230.

[131] *X & Y* § 27.

[132] IACtHR (Preliminary Objections, Merits, Reparations, and Costs), no. 12.353, *Escher and others, v. Brazil*, 2009, § 205-206.

One could think, for instance, of a toy robot that ends up making kids' information available in a manner that threatens their safe development. Would that case generate an obligation to mobilise criminal law to protect vulnerable children? The question remains open, certainly. And yet, ascertaining a right to a criminal response remains defendable.[134]

Equivalence triggers

This second assortment of triggers is more relevant to the kind of problems a robot will pose. As developed within the ECtHR in *X & Y*, it addresses situations where the victim's particular situation impedes accessing the criminal apparatus that is generally available to everyone.[135]

The point here is not the gravity but the state's choices. If it chooses to provide penal responses to certain behaviours, it should not leave specific categories of people unprotected because of their conditions. *KU* is a case in point. The case involved the publication of an advertisement on an internet dating site in the name of twelve-year old child.[136] Domestic laws sanctioned the conduct as an offence of misrepresentation. However, the specificities of the case —internet service providers privileges— impeded deploying the criminal apparatus.[137] Hence, the Court's found a violation of Article 8 (right to privacy) because the state failed to keep a framework capable of tackling that kind of cases.[138]

This second trigger is central for cases of robot harm. Emerging robotics and AI systems will likely commit the same old-age offences, like killing or injuries.[139] And, yet, like in *KU* and *X & Y*, how the law tackles these technologies might shield the human behind from liability. Think of a robotic surgeon that fails to cut some tissue appropriately and injures a patient.[140] If negligence in medical practice is generally considered an offence, why excluding the patient from the aegis of criminal law? Is it because the robot features impede attributing liability to the doctor, the programmers, or anyone behind its failure? If that were the case —the argument goes— the state would be failing to meet its human rights obligations.

[133] In general, the ICtHR's jurisprudence fails to draw clear triggers. On the contrary, it seems from its wording that any violation of human rights should lead to a criminal response (see Basch, (2008), 202-20.) The ICtHR's seminal case — *Velásquez Rodríguez* (1988) —- seems to imply that when it affirms that the *State has a legal duty to take reasonable steps to prevent human rights violations and to use the means at its disposal to carry out a serious investigation of violations committed within its jurisdiction, to identify those responsible, to impose the appropriate punishment and to ensure the victim adequate compensation* (§ 174).

[134] On this respect, see Law Reform Committee, (2021), 38-40.

[135] *X & Y* § 27.

[136] *KU v. Finland*, no. 2872/02, § 6-14, ECHR, 2008.

[137] Ibid., § 46.

[138] Ibid., § 50.

[139] Cf. Hayward and Maas, (2020), 214-18.

[140] The example is taken from O'Sullivan et al., (2019). See also Fosch-Villaronga et al., (2021).

### *ii. What do duties of redress demand?*

Now the triggers are delineated, it is convenient to turn to the question of what duties of redress involve. Imagine a state in any of the situations described above. That state will arguably need to do something to avoid a gap in attributing robot harm. But what is it exactly obliged to do? What do human rights demand of it? Is it enough to investigate, or should it also punish?

Undoubtedly, there is no right to obtain the prosecution or conviction of any person.[141] In other words, there is no *right to punishment*. That set a contrast between duties of redress, mainly developed under the ECtHR, and obligations to punish certain grave violations.[142] The later kind of right has found a broader acceptance, with developments within ICtHR.[143]

The kind of obligations that concern this thesis are different.[144] It is more accurate to say that states appear to bear a set of distinct positive obligations ranging from setting a framework to carrying out criminal investigations. In the recent case of *Volodina*, the ECtHR condensed three diverse commitments.[145] First, states must take reasonable measures in the face of a risk of ill-treatment that the authorities knew or ought to have known. Second, there is an "obligation to establish and apply in practice an adequate legal framework."[146] The third obligation demands conducting "an effective investigation."[147]

In *Beganović*, the Court added what can be considered a fourth obligation. According to the ECtHR, once an investigation reaches national courts, the case should be submitted to a "careful scrutiny" in a manner that the "deterrent effect of the judicial system is not undermined."[148] In this sense, states' obligation extends beyond the investigation to the proceedings before the judiciary. Even if the Court does not purport to discuss whether the judge applied

[141] *Ali and Ayşe Duran v. Turkey*, no. 42942/02, § 61, ECHR, 2008; *Beganović* § 77.

[142] See UNGA, (2005).

[143] See Basch, (2008). Cf. Velásquez Rodríguez (1988) § 176. As Kamber mentions, *the IACtHR imposes an obligation on the states to prosecute the perpetrators of human rights offences, who are frequently identified in its judgments, and thus it assumes certain functions of international criminal law* (2017), 176.).

[144] *Volodina* § 77.

[145] Ibid.

[146] See *Osman v. The United Kingdom* [GC], no. 23452/94, § 115, ECHR, 1998-VIII.

[147] See *X & Y* § 27. See also AComHPR (Communication*) Monim Elgak, Osman Hummeida and Amir Suliman v. Sudan*, no. 379/09, 2014, § 101, 138–141.

[148] *Beganović* § 77-78. Similarly, the AComHR has considered that states are obliged to "duly investigate, [and] prosecute the assailants and compensate the victims." (AComHPR (Communication) *Noah Kazingachire, John Chitsenga, Elias Chemvura andBatanai Hadzisi v. Zimbabwe*, no. 295/04, 2012, § 133)

national law correctly,[149] states must ensure that both the examinations of facts and the law's application are carried out with an adequate level of diligence.

It is plain that the whole spectrum of obligations will enter into play as smart robots are ushered. Already in 2010, for instance, some authors pointed out the need for a new type of robot-specific forensic science.[150] Despite their importance, this paper is mainly concerned with the obligation to design an adequate legal framework. Solving investigations hurdles or enabling courts to conduct sufficient scrutiny is undoubtedly relevant. However, it hinges on clarifying if the law is up to the challenge of attributing robot harm to a flesh and blood human — or a corporation. For what is the point of conducting an impeccable investigation if the law lacks the mechanisms for allocating responsibility?

Now, what is then the content of the obligation to set a legal framework? Is the Court rewriting the national criminal codes, as George P. Fletcher asserted? Or is it sufficient, for instance, to create separate robot offences and assign lower sanctions to humans behind?

*Volodina* is illustrative here. In that case, the Court found that the Russian legal framework failed to criminalise domestic violence. The reason is that it was not defined as a separate offense or an aggravating element.[151] Remarkably, the Court criticised both the requirement of a minimum threshold of severity and the demand that only assault of family members, committed for a second time within a year, could amount to a crime.[152] The Court indeed commented that demanding a minimum level of injury entails that psychological violence or economic abuse would fail to receive punishment.[153]

This latter case serves to highlight that not every kind of criminalisation meets the state obligation. Black letter law is not sufficient here. Instead, criminalisation requires "providing effective, proportionate and dissuasive sanctions."[154] Whether the state chooses to create new offences or reinterpret existing ones, it must provide *operative protection*.[155] A state will be falling short of its obligations if the framework has such a shape that some instances will remain out of it or be too tricky for victims to achieve some protection. Furthermore, such protection must bear some correspondence to the kind of offence it tackles. In general, thus, criminal law is there to shield certain rights, and it should be fit for purpose.

[149] Ibid., § 78.

[150] Sharkey et al., (2010).

[151] *Volodina* § 85.

[152] Ibid., § 81.

[153] Ibid.

[154] Ibid., § 78.

[155] See Mavronicola, (2020).

That affirmation raises more doubts than answers, nonetheless. How should the framework be shaped to be responsive to the required level of protection? What does it mean that a sanction needs to be dissuasive, proportionate, and effective? In order to answer these questions, it is necessary to look at the rationale behind criminal prohibitions. That is the next part's aim.

## C. A framework that is dissuasive, proportional, and effective

This section depicts a rationale for the ECtHR's duties of redress. Accepting the triggers and the obligation to set a framework, the question then is *why* criminal law an no other alternative –like compensation– receives the Court's favour. The Section tackles the question in three steps. The first step discusses alternative explanations (i). The next Sub-section (ii) discusses deterrence's role and, particularly, the importance of attributing criminal liability. Sub-section iii. zooms into the relevance for robotic harms.

### *i. Which theory of criminal punishment better explains duties of redress?*

The first question is why punishing and not, for instance, obliging someone to pay compensation?[156] The question is not purely theoretical. In a recent accident where an ADS failed to recognise a pedestrian and killed her, Uber —the vehicle owner and developer of the ADS— reached a settlement with the victim's family. That outcome felt unsatisfying for scholars like Ryan Calo, who admitted that a criminal case against Uber would have grappled "with what it means to build faulty technology."[157] What is so distinct from criminal law that it would have been a preferred alternative? What does it bring to the stage that is missing in civil lawsuits or other forms of compensation?

Theories go back and forth in answering that question. Some point to the good consequences that sanctions bring about. Criminal law contributes to aggregate well-being by preventing harm.[158] In the example above, attaching a punishment to whoever was behind the injury would somewhat avoid future injuries.  The question is then, *how* is it supposed to do so?

Here, again, theories vary. Some defend punishment as a means for locking up offenders and, thus, hindering them from continue harming.[159] Others point to deterrence, that is, not disabling the perpetrator, but making potential offenders less enthusiastic about offending.[160] There are two ways of dissuading likely-to-be offenders. The first way, called specific deter-

[156] Similarly, see Judge Kalaydjieva dissenting in *Söderman v. Sweden*, no. 5786/08, ECHR, 2013.

[157] Marshall, (2020).

[158] See Andenæs, (1974); Paternoster, (2010).

[159] Chalfin and McCrary, (2017).

[160] Ibid.

rence, entails discouraging whoever already committed a crime from further offending.[161] Here, specific deterrence would entail education or rehabilitating measures. The second way is general deterrence. Instead of focusing on the specific offender, it points to everyone in society.[162] It assumes that describing and attaching an undesirable consequence to it makes it unappealing for anyone to incur in such behaviour. Here, the actual punishment of an offender is nothing more than a confirmation that the state was serious about its intimidating message.[163]

In contrast to theories relying on the consequences of criminalisation and punishment, other authors point to the intrinsic value of castigating. What justifies attaching a penalty to certain conducts is not whatever good it might bring about. Whether it prevents crime or not is irrelevant. For these authors, criminalisation and punishment are reasonable if they give culpable actors what they deserve for their wrongdoing. Here retribution is criminal law’s primary rationale: states should punish because that is what offenders deserve. Dissuading future crime, or incapacitating offenders, plays no role.[164]

Which of the two provide a better explanation for duties of redress? Is it the fact that offenders deserve punishment for their wrongdoings? Or is criminal law’s ability to prevent right-undermining behaviours? ECtHR's words are a convenient starting point here. A framework that meets its standards is one providing “effective, proportionate and dissuasive sanctions.”[165]

The first implication of such wording is that it is its usefulness that justifies criminal law. Setting forth a criminal framework becomes an obligation if it achieves specific societal goals. That framing follows the gist of human right-based positive commitments. These obligations demand doing something to provide or achieve certain statuses that lead to the enjoyment of rights. And duties of redress belong to that family of commitment.[166] Thus, their justification hinges on their contribution to those statuses.

A corollary is that retribution theories cannot justify duties of redress. As pointed out above, these theories defend criminalisation and punishment irrespective of their social benefits. However, these social benefits are what justifies mobilising the criminal apparatus. Hence, an approach that sets aside them cannot explain why duties of redress are obligatory. Punish-

[161] Some authors defend the communicative aspect of criminal law. See, e.g., Tadros, (2007), 82. Consistently, Penny Crofts argues that badness or wickedness is communicated by criminal convictions (2013).

[162] See Farrell, (2015).

[163] See Pagallo and Quattrocolo, (2018), 401.

[164] For a summary of different positions, see Hart, (2008), 230-37.

[165] *Volodina* § 78.

[166] Lemmens and Courtoy, (2020).

ment, as what offenders deserve, cannot have the last word. If criminalising is part of a human rights obligation, it is because it contributes to the enjoyment of right-protected statuses.

Now, among the theories relying upon the usefulness of criminalisation and sanctions, general deterrence offers the best explanation—at least when it comes to the obligation to set forth a legal framework. Incapacitating an individual, or treating offenders, are not primarily a matter of how laws are drafted. Instead, they inform the obligations to prevent specific harms[167] and to investigate and prosecute offenders.[168] In contrast, what is at stake in the obligation to set a framework is general deterrence. A framework that includes effective, proportionate, and dissuasive sanctions has to be capable of persuading all the members of society to avoid committing offences.[169]

### *ii. What is criminal deterrence's specific contribution?*

Sub-section i. yields the following question: what is the specific contribution of criminal law to that deterrence? Why is paying compensation not enough? Imagine a high compensation or a lengthy procedure. Would those alternatives not suffice to disincentivise companies and individuals from tapping into technology or failing to ingrain safety safeguards. After all, off-the-court solutions are already common currency in the ambit of robotics and law.

When Knightscope's 300-pound security robot ran over a toddler, the company responded with a public apology.[170] Similarly, Tesla conducted its own investigation —and tweeted its results— to demonstrate that its autopilot system was not activated when one of its cars crashed and killed two passengers with no driver behind the wheel.[171] Reports of settlements, with companies paying compensation to injured victims, are also growingly frequent. Think of the USD 67 million that Intuitive set aside to settle claims for injuries that its surgical robot caused.[172]

Why is neither of those alternatives sufficient? What is unique of criminal deterrence when a high compensation or the prospect of publicly apologising might end up disincentivising harmful behaviours?  What does criminal law bring to the menu of alternatives, such that its deployment becomes a human rights-based obligation?

[167] See *A*.

[168] *Beganović*.

[169] Cf. Lemmens and Courtoy, (2020).

[170] Knightscope Issues Field Incident Report, (2016).

[171] Tesla: Elon Musk Suggests Autopilot Not to Blame for Fatal Crash, (2021).

[172] Compton, (2021).

The threat of a punishment is not the answer. Criminal law's deterrent power needs more than the allocation of a sanction.[173] Contrary to common assumptions, there is consistent research that most people base their decisions not on actual knowledge of legal sanctions but on notions of how the law ought to be.[174] It is thus the way criminal law's distribute criminal liability what counts.[175]

Indeed, the way criminal law distributes liability expresses the blameworthiness of certain behaviours and, in so doing, it guides society's conception of what is wrong.[176] Hence, the importance of fairly establishing a link with the offender, so as to tell her and others that what she did deserves blame. In this sense, criminal law's specific contribution is its "persuasive" power: it shields the enjoyment of rights by convincing individuals to abide by certain standards.[177]

Criminal attribution is thus different from civil law and other systems of attribution. Torts, for instance, are about allocating fixed losses. No one would dare to claim that they are only necessary whenever someone deserves moral condemnation.[178] When someone is condemned to pay compensation, what is at play is a distribution of economic losses that has nothing to do with the blameworthiness of the one called to pay. That person might not even play a role in that event leading to compensation.

Criminal law, conversely, stigmatises offenders. Someone is deemed criminally liable because whatever she did is not acceptable. Criminal law is the sole mechanism capable of bringing stigma, and with that ability, it offers a powerful and unique tool of behavioural control. Hence the practical importance of singling out the person *who did it* and speak to her (and to others about her).[179]

Another way to put it is departing from criminal law as performing "normative reconstruction."[180] On the one hand, shared normative ideas, practices, and institutions are part of what constitutes and sustains social life. Indeed, to be a society, every group requires a measure of solidarity around an embodied ethical life. On the other hand, crimes are communicative attacks on that ethical life: they threaten social solidarity by undermining the ideas, practices,

[173] Robinson, (2008), 175-212.

[174] A piece that has aged well, see Tyler, (1990). See also: Robinson, (2013); Robinson and Darley, (2007).

[175] Robinson, (2008), 175-212.

[176] Ibid.

[177] Ibid., 187.

[178] See Dyson, (2014).

[179] See Chapter 4.

[180] Kleinfeld, (2016).

and institutions at the foundation of social solidarity. In this scenario, criminalisation and punishment restitch the social order of values that an offence challenged.

Those theories not only have the advantage of clarifying the specific role of criminal law. They also provide an appropriate explanation of human rights jurisprudence. Why is it that duties of redress are only attached to particularly repulsive behaviours, like life-endangering measures? Why have not courts and treaties brought in statistics on the deterrent effect of certain rules? The best explanation seems that they are tracking common understandings of justice. That is why victim-specific gaps whenever the state has chosen a criminal law alternative give rise to an obligation to patch it; that is why severity, intention, and the victim's vulnerability are reasons to make criminal responses mandatory.

### *iii. Why is it important to solve failures of attributing robot harm?*

Both perspectives point to the importance of solving the hurdles that attributing robot harm might present. Somewhat paradoxically, indeed, the only way criminal law can fulfil its function —communicate or persuade in favour of the validity of social values— is by tracking a community's shared intuitions of what is fair and what deserves punishment. Liability rules deviating from those intuitions —by, for example, shielding from punishment those deemed responsible— ultimately undermine the law's credibility.[181]

In this sense, there is a consistent body of empirical research pointing to how criminal law fails to guide if it leaves unpunished conduct that the community deems morally condemnable.[182] Livermore and Meehl put it bluntly in the following extract:

> *just as the institution of law may be brought into disrepute by too easy attribution of criminality in situations where the label criminal is generally thought inappropriate, so also may the institution be undercut if it releases as noncriminal those society believes should be punished.*[183]

Now, turning to robot harm, how do attribution gaps impact the rationale of mobilising criminal law? Are failures of attribution a problem worthy of attention if criminal law's functions deserve to be safeguarded? The answer is yes. It is at least plausible that attribution failures will lead to undermining criminal law's persuasive or (re-) constructing role.

[181] Robinson, (2008), 176-88; Robinson and Darley, (2007), 18-31.

[182] French points to three mental states leading to vigilantism, that is groups of citizens enforcing law. The first of them is a sense of being deprived of justice (2001), 6.) Greene in turn uses an experiment to evidence how people come to disregard the law as the learn of cases they consider unfair (2003).

[183] Livermore and Meehl, (1966), 792.

The reason is that failures of attribution will undermine criminal law's persuasive activity—at least in the realm of technology.  Early empirical research shows that people assign liability to various entities behind the robot in an equative manner.[184] A failure to attribute robots' behaviour will arguably frustrate those expectations, leading to the deterioration of the law's persuasive role. And as robotics and AI become more and more ubiquitous, it is likely that such a decay will permeate beyond the specific context where those technologies are being introduced.

Some scholars already point to the problems of that deterioration. John Danaher insists that failures to apportion liability will likely lead to scapegoating, where, in light of those failures, society will target specific individuals or organisations with informal reactions.[185] Arguing in favour of directly castigating robots, Christina Mulligan has further defended the idea that robot punishment advances "the creation of psychological satisfaction in robots' victims."[186] If law's persuasive role is taken seriously, lack of psychological satisfaction will lead to criminal law losing its protective bite and, thus, the safeguard against right-threatening behaviours will end up deflated.

In any case, the failure will send the message that the social fabric is losing its ground to the introduction of AI and robotics. Such a message will undoubtedly invite developers, deployers, and whoever is behind the robot to disregard safety or security measures. That upshot is not at all imaginary, indeed. There are already voices arguing that as robots bring an overall reduction of deaths and injuries, it is justified to allow some erosion of (criminal) responsibility.[187]

The problem with such an approach is that, in the absence of a better explanation, the motivation to reduce harm partly stems from accepting social norms. What would be the fate of those norms if using a robotic system protects those behind it from liability? Would it entail a decay in safety, security, and other standards that make these devices reliable? That result is, at least, not unimaginable: an erosion of criminal liability might bring in the normalisation of offences jeopardising life, privacy, and personal integrity.[188]

Even assuming that safety norms would remain untouched, leaving harm caused by robotics outside the scope of criminal law is problematic. As Mireille Hildebrandt points out, it could

[184] Lima et al., (2020); Lima et al., (2020).

[185] Danaher, (2016).

[186] Mulligan, (2018).

[187] Santoni de Sio and Mecacci, (2021), 15.

[188] See Pagallo, (2013). As Helen Nissenbaum argues, "maintaining clear lines of accountability means that in the event of harm through failure, we have a reasonable starting point for assigning just punishment as well as, where necessary, compensation for victims." (1996), 2.

create a market for such technologies outside the censure of criminal law, leaving their regulation to whoever can afford the risk of tort liability.[189] That result is not satisfactory. At least not if the rights to life and privacy and the prohibition of ill-treatment are at play. Under their current configuration, it is unacceptable to leave some victims outside the protective realm of criminal law just because robots reduce the overall probability of accidents. For no overall benefit has trumped, nor there is a reason why it should, the right that potential victims have to a degree of protection by the state.

Overall, giving away with criminal law brings several problems. On the one hand, it is likely to undermine the overall protection of human rights. One persuasion tool away, and behaviours that threaten those rights might become common currency. Relatedly, it will disincentivise the embedment of safety and security procedures. If those behind the robot can harm without any blame, how could they have a reason to make it safer? Finally, it might leave aside some victims from the state's protective aegis, hence contravening the obligation to redress offences against the rights to life and privacy and the prohibition of ill-treatment.

## D. Conclusion

This chapter started with a question: why failing to attribute robotic harm would contravene human rights standards? The answer and first take away is that some rights-protected statuses – i.e., life, privacy and prohibition of ill-treatments— demand setting forth a criminal law framework capable of fairly singling out blameworthy individuals.

Arriving at that answer required various steps. Section A demonstrated that the right to a remedy, or a vague sense of criminal law's protective bite, does not offer an adequate response. Section B thus depicted an alternative approach based on ECtHR's duties of redress. It explained its triggers –gravity and equivalence—and described what states are obliged to do. Remarkably, it showed that those obligations include the duty to set forth a criminal law framework.

Section C turned to the reasons behind choosing criminal law as an obligatory response. It argued that criminal law's distinctive role is to persuade people into certain behaviours or provide a counter-message to the offender. Hence, the need to both track shared intuitions of justice and single out the individual who deserves blame. Such an approach not only has theoretical grounds, it also provides a fitting explanation of human rights-based obligations of redress.

[189] Hildebrandt, (2008), 169.

# 3 The atribution gap: prevailing positions

The previous Chapter sketched a human rights obligation to set forth a criminal law framework. Now it is time to look closer at the fate of such an obligation in a world where people share their space with robots. Indeed, the previous Chapter pointed somewhat vaguely to "attribution failures" as a problem of these technologies. That assertion begged the following questions: what are those "gaps," and why do robots bring in such a prospect?

Andreas Matthias is considered the first to introduce such an idea in 2004.[190] His point was that certain robots thwart the conditions for allocating responsibility to programmers, designers, users, or whoever is behind the machine. Robots might cause harm, and yet, because of their features, it is impossible to locate a suitable candidate to answer. Matthias's formulation of "attribution gaps" has been quite influential in early accounts of the challenges these technologies might pose to criminal liability. Particularly, Hildebrandt's seminal work develops them to argue that criminal liability would crumble as robots and AI enter into the scene.[191]

This chapter aims at understanding —from the perspective of emerging scholarship and reports— how it is that robots pose an attribution gap. It suggests classifying the gap into three categories. To this end, Section B introduces the ideas of attribution gap as a (i.) "complexity problem," (ii.) "unpredictability problem," and (iii.) "autonomy problem." That account sets the stage for the following chapter, where the piece presents some objections to those theories and proposes reframing the puzzle.

However, before engaging with existing approaches, it is necessary to ask what is an attribution gap. Without some definitional clues, it becomes easy to confuse such gaps with other problems. For the sake of analytical clarity, thus, Section A sketches a working definition.

## A. Defining attribution gaps

An appropriate concept of attribution gaps should have two functions. Firstly, it needs to be comprehensive enough to reflect the stakes of emerging debates. At the same time, however, it should be well-cut to avoid mixing up those debates with other concerns. For instance, victims' difficulties in retrieving AI software as evidence in criminal processes are often mixed up with attribution gaps.[192] However, both problems are different. Indeed, one thing is to say that AI complicates *demonstrating* that someone is liable, and another is to say that such an allocation of liability is impossible from the outset. The first is a problem of gathering evi-

[190] Matthias, (2004). Cf. Karnow, (1996).

[191] Hildebrandt, (2008), 167.

[192] See e.g., Abbott's Enforcement Problems (2020), 114.

dence to demonstrate that someone is liable, and hence demands fetching tailored solutions to facilitate victims' access to information.[193] The second, on the contrary, involves the allocation of responsibility itself. It asks not whether parties in a process can gather evidence to blame someone but whether that blame is even thinkable.

Köhler et al., propose characterising that latter set of issues as a "normative mismatch." Individuals and society at large typically seek to allocate responsibility. Since robotics and AI make it impossible to find a suitable candidate, they yield a discrepancy between social expectations and the reality that no one can receive it.[194]

However, claiming that attribution gaps are a normative mismatch begs a further question: what kind of norms count for such a mismatch? Which kind of standard is that robots frustrate? For Danaher, such benchmarks stem from the human inclination to identify those deserving punishment. [195] The predisposition to find someone who "pays" for the harm yields an expectation that such a person should be singled out. Under this account, robots and AI frustrate that tendency to the extent that they make it intractable to find the person deserving punishment.

Andreas Matthias, in turn, has moral norms in mind when he addresses the problem of attributing robot's misbehaviour.[196] His view is that some ethical standards (?) demand establishing a causal link between an outcome and a person who *voluntarily* decided to engage in harmful conduct. And robots frustrate that inclination.

It is submitted that human inclinations or moral expectations cannot be the key to identify criminal law norms. Even if related, the latter is a separate system and has its own liability rules. Those rules are what count to establish the mismatch. Hence, the definition's purpose is to properly frame the kind of vision of criminal law that the prevailing position coaches. Which features of the latter they have in mind when they claim that robots frustrate criminal liability? That is the question. In this vein, attribution gaps will arise whenever:

> *(1) it seems fitting to hold some person(s) to account for some φ to some degree D. (2) In such situations, either (2.1) there is no candidate who it is fitting to hold to account for φ or (2.2) there are candidates who appear accountable for φ, but the extent to*

[193] See Chengeta, (2020), 7, 10-11.

[194] Köhler et al., (2017), 54.

[195] Danaher, (2016).

[196] Matthias, (2004).

> *which it is, according to criminal law regimes, fitting to hold them individually to account does not match D.*[197]

(1) Points to the conditions to hold someone accountable for some φ. One cannot say that robots pose an attribution gap without presupposing some requirements for saying that someone is liable. Put differently, there is no mismatch if one cannot describe the conditions for a suitable match. Describing the conditions that they have in mind is thus necessary to know why robots frustrate them. That is precisely what this part of the definition encapsulates.[198]

(2) Shifts attention to the second requirement of the mismatch: assuming that the conditions in (1) are ά, έ, and ί, failing to meet one of them would yield the result that there is no fitting candidate for liability. Here it is important to stress what the word "suitable" signifies. It does not mean that no one will respond. Instead, the point is that, even if someone is found liable, that someone is not an *appropriate* candidate.[199]

That leads to the final point (2.2.): attribution of liability does not match the degree to which that person should receive blame. It leaves room for claims that, even if someone receives blame for the robot's behaviour, because of that intervention, the regulatory response does not meet the degree it would in other circumstances.

### B. Amid autonomy, unpredictability, and complexity: the criminal liability gap

This Sections uses that concept to engage with the different theories. It asks (1) what the notion of responsibility at stake is and what does it demand? and why do robots (2.1.) frustrate attributing liability to someone or (2.2.) or attributing it to an acceptable degree? Sub-section i. addresses who depart from the robot's complexity. In turn, Sub-sections ii. and iii. focus on unpredictability and autonomy respectively.

#### *i. Attribution gaps as a "complexity problem"*

Consider an example. On the evening of 18 March 2018, an automated driving system (ADS) led a Volvo XC90 to run over and kill a crossing pedestrian.[200] Under criminal law, it would be suitable to hold liable those who caused the pedestrian's death and had such a level of

[197] Köhler, Roughley, and Sauer, (2017).

[198] See Chapter 4.

[199] The point, Chapter 4 argues, remains underdeveloped.

[200] See National Transportation Safety Board, (2020).

knowledge and control over the outcome that they could have done otherwise.[201] Who could meet those requirements here?

It seems natural to start with the ADS programmer. However, that functionality entailed no less than three systems, each with various software and hardware components.[202] Different programmers were involved, including a third-party company, in its development.[203] On top of that, someone within the company decided to override the car's factory-built braking system—which would have prevented or mitigated the accident.[204]

Who among them made the decision leading to the crash? Who had such a level of control over that harmful outcome?

For some authors —including UN's former Special Rapporteur on extrajudicial, summary, or arbitrary executions— that question is intractable.[205] Robots and AI systems form a complex imbroglio of human collaborators and artificial systems such that no one can be said to cause harm. Criminal law demands holding to account those who caused and had control over the outcome. However, robots are so complex that either there is no unique cause or no one fully controls the result. Hence, there is no fitting candidate to hold to account whenever things go wrong.

Abbott calls these problems "practical irreducibility" and "legal irreducibility."[206] In his opinion, it would be "impractical" to reduce robotic conduct to individual human action because of "the number of people involved, the difficulty in determining how they contributed to the AI's design, or because they were active far away or long ago." Even if possible, it might be unfair for reasons of "criminalisation policy" to blame a person.[207] The reason is that the risks that each careless individual might generate in such a mess-up are not substantial enough to receive a criminal response.[208] In a similar vein, Pagallo argues that robots bring in a complex network such that there would be a "failure of causation," that is, harm whose origin remains obscure and, thus, hinders criminal attribution.[209]

[201] See, e.g., Hildebrandt, (2008), 168.

[202] National Transportation Safety Board, (2019), 5-6, 8-11.

[203] Ibid., 9.

[204] Ibid., 39-41.

[205] UNGA, (2013). para. 14-16.

[206] Abbott, (2020), 114-15.

[207] Ibid.

[208] Ibid., 114.

[209] Pagallo, (2013), 73.

Mireille Hildebrandt, in turn, uses the concept of "distributed intelligence."[210] Her point is that it is impossible to trace intelligence emerging in a network of persons and machines to one or more of its constituent parts. Returning to the example above, Hildebrandt would argue that the car's decision-making capabilities that led to the accident did not stem from an individual decision. On the contrary, the device brings in several layers of human collaboration, service providers, mechanical parts, and software applications. Thus, intelligence ends up diffused within an entanglement of persons and machines. The responsibility might be distributed beyond measure, such that no one can respond.[211]

Both accounts echo the problem of "many hands." Any complex array of individuals and processes, not only robots, make it challenging to identify who is to blame.[212] However, for scholars, robots bring in an additional element to the problem. As Susanne Beck argues, robotics is not only a collaborative project between persons.[213] It also entails transferring some responsibility to the machine, such that the mess-up becomes even more complex. In the same line, Ugo Pagallo locates the risk of attribution failure in the "intricacy of the interaction between humans and computers."[214]

The car crash example above illustrates the point. If one were to attribute responsibility, one could not solely consider the distribution of roles among humans. It is also necessary to factor in the software and hardware components and the decision-making capabilities embedded in the machine. All in all, it was the obstacles recognition system —and not the input of any individual— which failed to recognise the pedestrian. In the views of Beck and Pagallo, those machine components create a new version of the "problem of many hands;" one that muddles humans and machines in an intractable entanglement.

In sum, failures of attribution as a problem of complexity rests on the idea that (1) it seems fitting to hold accountable those who caused the robot harm and had enough knowledge and control over the outcome to avoid it. However, in the opinion of many scholars (2.1/2.2.), no candidate is fitting to hold to account fully. The reason: no one had enough knowledge and control in the complex mess of machine elements, software, and humans that make up a robot.

[210] Hildebrandt, (2008), 168.

[211] Ibid.

[212] Nissenbaum, (1996).

[213] Beck, (2016), 141.

[214] Pagallo and Quattrocolo, (2018), 386. Similarly, the Singapore Academy of Law Reform Committee argues that such a complexity raises the problem of "which aspect of the RAI [robotic and AI]system factually caused it to act the way it did (resulting in harm);" "which party (or parties) – be that the system manufacturer, the system owner, a component manufacturer, or a software developer – was responsible for that aspect," and "whether that party could have foreseen or mitigated the harm." (Law Reform Committee, (2021), 1-2.

### *ii. Attribution gaps as an "unpredictability problem"*

#### General overview

Consider the following example. In 2016 Microsoft launched Tay, a chatbot programmed to interact with users on Twitter. The digital agent's conversation pattern remained undetermined to a specific —still unknown— extent. The agent was instead supposed to "learn" — that is, to modify its programming— through its interaction with other users.[215] The bot, however, ended up tweeting misogynous and offensive comments. According to a later blogpost, that behaviour was not programmed into the agent.[216] Nor its developers and deployers could foresee that it would end up posting those tweets.[217]

Assuming that Tay's behaviour is a criminal wrongdoing, to whom could it be attributed? For some authors, the only suitable candidate is the one who could at least foresee the harm. Apart from causation, these authors posit a demand that the perpetrator had the knowledge and the capability to overturn the harmful outcome.[218] Those who cannot predict the ensuing behaviour cannot know and control it, for knowing presupposes being able to anticipate the consequences of one's behaviour. Now, robots impede predicting those results, as their behaviour is inherently unforeseeable. And since that anticipation is necessary to say that someone had sufficient knowledge and attribute liability, robots make it impossible to blame anyone entirely.[219]

Those circumstances generate a liability gap. Tay engages in harmful behaviour such that it would be fitting to accuse someone and, yet, because of its features, there is no suitable "someone" to carry the burden of blame. "In such cases, there would be no human to hold directly responsible for the decision to attack"[220] because such a decision could not be foreseen.

The question of why robots are unforeseeable has received various answers, nonetheless. For authors like Mireille Hildebrandt, that lack of foreseeability is a problem of explainability.[221] Others instead point to the robot's ability to exhibit emergent patterns, that is, behaviour that neither of its parts could explain.

[215] Lee, (2016).

[216] Vincent, (2016).

[217] Lee, (2016).

[218] Ibid.

[219] Ibid.

[220] HRW and IHRC, (2014).

[221] *Systems that learn in unpredictable ways and generate solutions that even computer engineers cannot explain seem to have acquired a new type of agency, not fully determined by the intentions of the designers.* (Hildebrandt, (2008), 169.).

### Unforeseeability as a case of the obscurity of decision-making processes

Going back to Mireille Hildebrandt's position, the fact that some robots exhibit ML capabilities implies that the process to map the data and produce an output remains obscure. Engineers might explain the data that the device acquired and might be able to comprehend the output. However, they cannot understand *why* that, and no other, was the outcome.[222] Similarly, Jansen and Brey argue that the "lack of transparency in AI systems," like those exhibiting some form of ML, makes it harder to ascribe responsibility to any individual for the harm these devices might cause.[223]

Fully understanding the argument leads back to their account of responsibility. For a candidate to attract liability, she must at least anticipate the result. Now —the argument goes— such anticipation requires understanding the process leading to that result and, to a certain extent, being able to explain it. Since AI systems are so obscure that no one can understand why a specific outcome came to be, they become unpredictable and, hence, criminal liability ends up frustrated.

Consider a further example. On the afternoon of 7 July 2016, a conical, bulky, AI-powered K5 robot was roaming with its four internal cameras at a mall. It then ran over a child scraping its legs and causing bruises. According to the company's explanation, the machine sensors "registered no vibration alert," and the robot did not act as it should have when encountering an obstacle.[224] Why that was the case, however, remains a mystery. As the company report, each K5 hosts nearly 30 sensors and a software stack, such that it can sense its environment from less than an inch away to over 91 meters. The only thing the company did was considering the accident as a "freakish" one. However, they could not explain what went wrong.[225]

Anticipating the consequences of one's actions is the reason why people are deemed liable. How could it be possible to anticipate those consequences whenever the processes leading to it remain obscure? It is precisely the inability to understand the thinking process that renders robot's behaviour unpredictable.

### Unforeseeability as a case of emergence

Ugo Pagallo affirmed that "robots stress common standpoints of what can be considered as a natural or probable consequence of a certain behaviour."[226] He seems to mean that robot's

[222] Ibid.

[223] SIENNA project, (2019), 74.

[224] Vincent, (2016).

[225] Holmes, (2020).

[226] Pagallo, (2013), 49.

complexity makes their actions unpredictable for whoever designed its parts. Robots are a mixture of different hardware and software parts, with sensors and cloud services. Their parts interact, and the results they bring in do not depend on those parts but their interplay. And that interplay remains beyond the control of whoever developed or used the robot.[227]

The idea that responsibility entails anticipating harmful results is also at stake here. Suitable candidates are those who *could* predict the upshots of their behaviour or those of their tools. Since developers cannot predict a robot's behaviour, they should not receive criminal blame. Several reasons are at play here.[228]

First, robots react to data that they gather through sensors or in interaction with their environment. As long as a developer does not feed up that data, she cannot predict how the robot will behave.[229] Second, some robots have many parts, and those parts might interact between themselves in surprising manners. Moreover, some robots are open because several applications might end up embodied in the same physical platform. In those cases, whoever developed the physical platform cannot foresee all the harms that its creation might end up causing—nor will it be possible for software developers to fully appreciate the interplay of the digital application and the physical layer.[230]

Third, some robots modify their programming as they interact with their environment. Think of ML techniques, where the robot "learns" to identify patterns in data. As with the interaction of many parts, a robot that changes its programming might end up exhibiting behaviour outside its initial parameters.[231]

How could a developer anticipate a robot's behaviour if the ingrained programme changes as it interacts with the environment? How could those intervening in creating an open platform predict that it becomes harmful with certain applications? Or how could those installing the application comprehend the platform's reaction? For some, it is unfeasible. And since unanticipated behaviour does not attract criminal liability, developers will not receive any blame.

If developers are not to blame, who else could be a suitable candidate? Could it be a user who did not intervene in the design of the harmful tool? Or could it be an innocent third party who ended up interacting with the robot? Imagine that, in the accident leading to Knightscope's robot, one decides to blame the parents for letting the child interact with a robot who ended up causing harm. The authors do not address the point explicitly. However, one can make them

[227] Calo, (2015), 538-45.

[228] Cf. for the IHL literature, Chengeta, (2020), 3.

[229] Cf. Pagallo, (2013), 49.

[230] Ibid. See McAllister, (2017), 2533-36.

[231] These are the main focus in Matthias, (2004).

to say that the robot's behaviour is as unpredictable for users as it is for developers, since it is irrational to expect users to predict the behaviour of a tool they had no role in devising.[232]

Go back to the example of Microsoft Tay. It absorbed users' tweets and learned to interact. All the features of the robot were designed to make it a friendly conversation partner. However, these features rendered quite different results. Instead of a pleasant conversation partner, Tay ended up as an insulting misogynous agent.[233] Authors like Ryan Calo[234] would argue that those who developed Tay could not attract blame, as they did not foresee such interaction with the environment. As Microsoft pointed out, the company had already launched an agent in another social network, and the agent did not end up insulting other people.[235]

Could the users receive blame? The Guardian picked out an example when Tay was having an unremarkable conversation with one user. The user asked: "is Ricky Gervais an atheist?" and Tay replied saying: "ricky gervais learned totalitarianism from adolf hitler, the inventor of atheism." Nothing in that user's behaviour reveals that it could predict such an answer. Assuming that it entails an offence, thus, it seems unreasonable to blame @TheBigBrebowski for such a surprising response.[236]

### *iii. Attribution gaps as an "autonomy problem"*

The last account points to autonomy as an explanation of attribution gaps. It comes from the realisation that criminal systems do not call persons to account for the behaviour of other agents that are also accountable.[237] Conversely, vicarious liability, where an agent answers for whatever others do, goes against some of the central tenets of criminal law.[238] The latter aims at shaming individuals for their behaviour as a deterrent. How could that shame be effective if it was not the individual but another agent who engaged in the behaviour? For some authors, the combination of criminal law principles and robot autonomy poses an intractable obstacle for attribution.

That is the type of problem that future AI systems will pose. Indeed, some of these systems will come to exhibit such a level of autonomy that they might meet all the requirements for

[232] Pagallo's following words are as relevant for developers as they are for users: "Robots stress common standpoints on what can be considered as a natural or probable consequence of a certain behaviour." (2013), 49.).

[233] Vincent, (2016).

[234] Calo, (2015), 538-45.

[235] Lee, (2016).

[236] Hunt, (2016).

[237] See Chapter 4.

[238] Arguing that it is possible within some conditions, Diamantis, (2021).

being liable.[239] Hence, they will block the path for attributing liability to the human behind. As Simmler and Markwalder argue, the robot under this hypothesis has developed its "own momentum" due to its artificial intelligence. That momentum cannot be traced back to a single programming situation. Instead, it poses a problem similar to the debate on free will.[240] Indeed, it is impossible to trace the robot's behaviour back to one of its determinants. The device is so autonomous that no one can tell that she defined the device's behaviour.

Lagoia and Sartor explain how robots might exhibit the kind of autonomy that criminal law demands.[241] Using Weyns et al.'s model of perception, it seems possible for robots to perceive the environment and generate interpretative models.[242] Bratman's framework, in turn, serves to demonstrate that artificial agents can form intention by committing to certain goals, even if those goals are not freely chosen by the device but by a programmer or a human user.[243] In this sense, a robotic surgeon who chooses the speed up the cutting can be said to have acted intentionally, even if it is a human operator who maps the area where the cut is to be done. [244]

Furthermore, some devices will be reason-responsive and exhibit the ability to act otherwise. All that is needed is the capacity to be aware of their own behaviour, and understand the impact of sanctions on their interests. [245] These devices present a situation similar to that of a partial psychopath, who is also accountable for her offences. It is thus no obstacle that the device cannot apprehend the moral wrongfulness of its acts.[246]

Many authors stress here that, to say that a robot is autonomous, they do not need to exhibit human-like capabilities.[247] In contrast, it is sufficient for them to comply with the conditions that criminal law demands: awareness of the environment and ability to commit to certain goals, along with self-representation of one's behaviour and interests. Achieving those capabilities are undoubtedly feasible and, once achieved, it will not be possible to attribute their behaviour to a human behind. Doing otherwise would entail infringing upon the tenet that criminal is strictly personal. As Gabriel Hallevy reasons, in this scenario, "there is no reason" to prevent the imposition of criminal liability upon the AI system itself.[248]

[239] Lagioia and Sartor, (2020).

[240] Simmler and Markwalder, (2018), 8.

[241] Lagioia and Sartor, (2020), 436-51.

[242] Weyns et al., (2004).

[243] Bratman, (1987).

[244] See Connor et al., (2020).

[245] Lagioia and Sartor, (2020), 448-51.

[246] Ibid.

[247] See Hu, (2019).

[248] Hallevy, (2010), 184. See also Hallevy, (2015), 47-184.

Determining whether that should be the case or not is not the purpose of this Section. Here it is sufficient to note that such affirmations of autonomy entail that criminal liability cannot be traced back to a human agent, thus presenting the prospect of an attribution gap. Indeed, (1) it is fitting to hold accountable *only* those agents who committed to the behaviour being aware of the sanctions and its wrongfulness, such that the behaviour of others cannot attract criminal liability. Now, (2) in certain situations, AI will exhibit those traits such that (2.1.) no one — except for the artificial agent— will be a fitting candidate to hold to account. Therefore, an attribution gap will be present.

## C. Conclusion

The Chapter set to the task of understanding —from the perspective of emerging scholarship and reports— how is it that criminal law's tenets will crumble with the introduction of robotics and AI. It firstly defined attribution gaps as a normative mismatch between criminal law's conception of responsibility and the reality robots bring in. Then it employed such a definition to clarify that, for scholars and organisations, robots present a gap because:

> (1) Under their vision of criminal law, someone is accountable provided that she (*i.*) caused a particular outcome, (*ii.*) with knowledge and control over its development, and (*iii.*) with the ability to understand the implications of her behaviour.
>
> (2) Whenever a robot harms it is impossible to find a suitable candidate because either (*i.*) the latter's behaviour leads to a complex setting where no one can be said to have caused the outcome; (*ii.*) or because it is so unpredictable that no one could have known the possibility of such an outcome; (*iii.*) or because the robot itself fulfils the conditions for being a suitable candidate, such that the human behind cannot be the subject of criminal blame anymore.
>
> Such a situation leads to a scenario where (2.1) either no one responds, or (2.2.) no one responds to the degree that it would have if no robot were present.

"Attribution gap theories," as this thesis coins them, thus presuppose a particular vision of criminal law and of what attribution failures mean. The next Chapter contests both views and proposes a new method for disentangling criminal accountability and keeping the bite of duties of redress.

# 4 Technologically blurred accountability: A model to disentangle attribution gaps

The previous Chapter presented an unpromising panorama for human rights. The digital age is rushing in technologies that —according to many— will frustrate the attribution of criminal liability. The entrance of robots into society seems to come at the high costs of eroding and obscuring criminal accountability and, with it, the sheer feasibility of duties of redress.

How to solve that problem? Some voices suggest that shifting liability to the robot might be the solution,[249] whereas others recommend restricting the introduction of technology.[250] On the contrary, the thesis submits that it is convenient to take a step back and refine the prevailing understanding of attribution gaps. This Chapter introduces an alternative model to solve those cases of unpredictable, complex and autonomous machines. On the one hand, the model demonstrates that robots do not blur attribution. On the other, it serves to highlight two instances that, despite their potential to cripple criminal accountability, have remained thus far ignored.

The first step to develop that model demands contesting the vision of criminal liability that the prevailing position couches. As shown in the previous Chapter, attribution gap theories seem to assume a particular notion of criminal liability. In their opinion, someone is accountable provided that she (*i.*) caused a particular outcome, (*ii.*) with knowledge and control over its development, and (*iii.*) with the ability to understand the implications of her behaviour. Section A argues that such a model is unnecessarily narrow. Tapping on human right standards, mainly developed within the ECtHR's jurisprudence, it maintains that it is also conceivable to deem liable an unaware subject precisely for failing to realise the risks she was creating through a robot. Section B turns to *the how*. In what way could one say that the human behind the machine is criminally accountable for what the latter does? It suggests three baseline conditions for saying that an unaware person is criminally liable whenever the machine misbehaves.

The second step has to do with the last requirement of the definition. Attribution gap theories present only two scenarios: either no one responds, or no one responds to the degree that they would. Such an approach fails to capture what is specific of attribution gaps. As Section C argues, the problem is not whether someone is liable, or not, but whether deeming that person liable meets the human rights prerogatives of both defendants and victims. Departing from

[249] Beck, (2016), 141-42.

[250] Hildebrandt, (2008), 178.

duties to redress and defendant's right to a personalised blame, the Section defines three scenarios of *technologically blurred attribution*.

### A. Looking at the person behind and broadening the vision of criminal liability

More than with the tools they use, criminal law is concerned with the way people make decisions and what those decisions tell of an individual's character.[251] Attribution gap theories put the stress on the machine, of whom they asks if it is unpredictable, autonomous and complex. This thesis suggests an alternative approach. Instead of starting from the machine, it starts by describing the person "behind" it. That is the goal Sub-section i.

Considering such a description, Sub-section ii. defends the need to broaden the first of the requirements that make up attribution gaps. It suggests that instead of portraying an accountability problem, the prevailing position's approach reflects a too narrow view of criminal liability; one that not only fails to account for many domestic laws, but that also falls short of grasping the scope of states' obligation to mobilise their criminal apparatus.

#### *i. The human "behind" the machine*

Criminal liability seeks to determine whether the agent deserves blameworthiness.[252] Thus, the question of why technology presents an attribution gap is not so much one of whether the machine was unpredictable or autonomous. On the contrary, it is one of what the agent did and what she had in her mind when she did it. And the main feature of that agent, as attribution gap theories depict her, is that she did not know that the –unpredictable, autonomous or complex— robot would end up harming.

Certainly, technologies are different. Mireille Hildebrandt's smart environment that orders poisoned bread erroneously delivered to a neighbour's house[253] has little to do with Pagallo's Robot Kleptomaniac —which steals a shop to get batteries. [254] And neither of them resembles the kind of LAWS used in law enforcement.[255] However, differences fade away as soon as one focuses on the "human behind."

Indeed, what these cases have in common is a "human behind" who fails to realise that the machine is exposing others to danger. The scenario is one of a robot who behaves in a manner that the defendant was not aware of. LAWSs, the smart environments, and a *Robot Kleptoma-*

[251] Stark, (2016), 179-86.

[252] See Chapter 2.

[253] Hildebrandt, (2008).

[254] Pagallo, (2013).

[255] HRW and IHRC, (2014).

*niac* are very different devices. However, if one shifts the focus from the machine to the person behind it, the result is a user, programmer, developer, or integrator who did not realise that the robot would end up harming as it did.

Claims of unpredictability, autonomy and complexity make no sense outside these scenarios. For it is self-defeating to say that the harm was unpredictable whenever the agent knows and intends to harm. Indeed, aiming at something presupposes anticipating it as a possible result. What if the machine is autonomous? Well, having an autonomous partner in a deliberate offence, even if that partner is artificial, does not exclude attribution. On the contrary, either the machine becomes a tool in the hands of the offender,[256] or the offender is jointly liable with the machine. [257]

Again, if one among the "many hands" introduces a line of code anticipating and aiming at harming someone, would it not be possible to single that "hand" out of the imbroglio, such that no problem of complexity arises? In fact, approaches like Abbott's legal irreducibility presuppose many unsubstantial contributions[258] As soon as a contribution aims at causing harm, it has the components to become substantial and then stands out within the network.[259] The point is even clearer when it comes to Hildebrandt's suggestions. Recall that the latter focuses on the impossibility to be aware of the outcome as one is part of a complex network of machines and agents.[260] Undoubtedly, that impossibility fades away as one of the nodes can realise the harmful outcome that the robot is bringing in.

Hence, a first step is to recognise the kind of affordance —i.e., possibilities of action— these devices introduce.[261] The technology allows causing harm without whoever is behind being aware of it. They thus introduce a not-knowing and not-intending subject. Yes, the robot might be complex or autonomous. But what is crucial to notice is that the human behind ends up "offloading" her knowledge about such a harm. LAWSs allow offloading the ability to select targets and, in some circumstances, engage them.[262] The smart environment allows de-

[256] See e.g., Simmler and Markwalder, (2018), 7..

[257] At most, if the robot is highly autonomous, these cases are comparable to situations of innocent agency—where the human behind uses a robot agent, the latter not knowing the situation nor understanding its own infraction. One could also think of cases of complicity—where the "human behind" joins forces with the artificial agent to commit an offence (see Hallevy, (2015); Hallevy, (2010).

[258] Abbott, (2020), 114.

[259] Certainly, the 'but for' test would demand more than just being aware. In general, that a cause is substantial means that the result would have not ensued had supressed such a cause. However, some writers also distinguish between causation and imputation, the latter adding that the specific risk created with an act materialises in the result. For an account, see Hart and Honoré, (1985), 389-94.

[260] Hildebrandt, (2008), 177-78.

[261] See Liu et al., (2020).

[262] See HRW and IHRC, (2015).

ferring the management of a house, including food purchases.[263] And Pagallo's Kleptomaniac permits delegating the planning and execution of robberies.[264]

Such a feature is behind the mismatch yielding an attribution gap. Recall that what makes it impossible to find a fitting candidate (2) is that criminal laws call for one who at least knew and had control over the harmful outcome (1.*ii.*) while retaining her ability to grasp its blameworthiness (1.*iii.*). The human "behind" the machine is someone who falls short of those conditions. That is what makes her an unfitting candidate. Had her known and controlled the robot's development and deployment, it would have been possible to call her to account. That is the gist of attribution gap theories.

### *ii. The narrow viewpoint of prevailing positions*

Now, the question is: which kind of attribution model can capture such a scenario where the person behind does not know the risks that the machine was creating? Recall that, as argued in Chapter 2, criminal law is about communicating blame to an offender. In contrast to other regimes that are concerned with the distribution of costs or the allocation of some benefits, criminal law is after the person who can be say to own the wrongdoing.[265] If an automated vehicle erroneously classifies and knocks downs a pedestrian, criminal law will not be content with saying that someone ─an insurer, for instance─ should bear the costs. It rather seeks to "speak" with whoever "owns" the harm to tell her and others that whatever she did is wrong and ought to be avoided.

The prevailing position would say that it is impossible to speak with her whenever there is an autonomous or unpredictable machine in the picture. In their view, the only circumstances where one could do so is when she at least knows and control the outcome. And, as the previous Sub-section showed, that is precisely what robots frustrate.

The thesis submits that such a view is too narrow. It sets a standard for liability that it is unnecessarily high. Certainly, blaming the not-knowing subject is a contested matter.[266] However, it is already a feature of many criminal laws that the careless, inadvertent subject, is also a

[263] Hildebrandt, (2008), 176-77.

[264] Pagallo, (2013), 53-54.

[265] Hruschka, (1986), 669-71.

[266] Some argue that "There is no moral difference between punishing for inadvertent negligence and punishing on the basis of strict liability, and the lack of a moral difference evidences itself in the inability to draw a distinction between strict liability and negligence on any basis other than arbitrary stipulation." (Alexander et al., (2009), 81-85; Cf. Husak, (2010).

target of criminal blame.[267] Furthermore, duties of redress demand mobilising the criminal apparatus not only when the defendant knew and controlled the injuries she caused to her victim. The HRCtte's, ICtHR's and, to an extent, the ECtHR's jurisprudence showed that those obligations also arise in cases where a state faces the harm that a negligent —and hence inadvertent— subject caused. [268]

It is thus necessary to broaden the view of criminal law that the prevailing position coaches. Far from reflecting a failure of accountability, such a notion might reflect a state who falls short of meeting its human rights obligation to set an adequate framework. The question now is: how to blame someone that lacks knowledge? How could states fulfil their obligation to set a framework that is dissuasive, proportional and effective if their target is someone who was not even aware of the harm she was causing through a machine? Refining that view is the next Section's goal.

## B. Blaming the not-knowing and not-intending agent behind the machine: an alternative vision of criminal liability

Is the fact that the person behind the machine does not know nor control it critical? Prevailing positions would say yes. As shown in the previous Section, they assume that attribution needs, at least, knowledge and control over the machine's deployment or development. How to blame someone that precisely lacked those features? This Section shows how it could be possible. It builds its position on the human right to a personalised blame, mainly developed within the ECtHR. Sub-section i. argues that the later only requires a "mental link;" one that is broad enough to accept blame not only for the choices one makes and controls, but for failing to make them.

However, that does not mean that states can go willy-nilly and deem anyone liable for robotic harm. That human rights standards do not require control does not mean that they do not require any mental link whatsoever. Sub-section ii. introduces three conditions that would allow to establish it in the circumstances of the unaware agent behind the machine.[269]

[267] Building a "Standard Account" of Common Law countries, see Stark, (2016). Comparing the English and the French system, see Spencer and Brajeux, (2010). For a description of the German, Spanish, Russian and British systems, see Heller and Dubber, (2011), 252-87, 414-54, 88-562.

[268] See Chapter 2, Section A, Sub-section i.

[269] Introducing the importance of liability for negligence and explaining some of the practicalities, see also Law Reform Committee, (2021), 29-30.

### i. *A human right to personalised blame*

#### The obligation to establish a connection

Declarations of criminal liability intend to ‘get personal’ in a way that the law’s other declarations —awards of civil damages, determinations of tax liabilities, etc.— do not. Such a personalized blame is not only gist of criminal culpability,[270] and an important component of duties of redress.[271] It is also a human rights obligation. One can arguably read such a demand in the rights to be presumed innocent[272] and on states’ obligation to define precisely by law all criminal offences.[273]

Among the bodies, the ECtHR is the one who has given its shape to such a demand —and hence its jurisprudence is the main focus of this Sub-section. Why is that the case if the Convention does not expressly oblige to get personal? What could be the content of such a right?

In the landmark *GIEM SRL and others*, the Court build that position on the basis of Articles 7 and 6 (2) of the Convention.[274] Understanding why is that the case requires first unpacking the content of Article 7.

Broadly speaking, the latter enshrines the prohibition to blame a person for offences that were not previously established in the law.[275] Be it a black-letter enactment or a judiciary’s jurisprudence, whether a prohibition is established in “the law” demands meeting two criteria.[276] The first of them is “accessibility.” It means that criminal prohibitions are public, that is, they are in an instrument which is at hand to its addressees.[277]

However, it is the second condition —“foreseeability”— that hallows the obligation to “get personal”[278] when it comes to imposing criminal blame. Foreseeability means that provisions must be clear enough for an individual to know “what acts and omissions will make him crim-

[270] Stark, (2016), 179-86.

[271] See Chapter 2.

[272] ICCPR, Art. 14 (2); ACHR, Art. 8 (2); ACHPR, Art. 7 (1); Convention, Art. 6 (2).

[273] ICCPR, Art. 15 (1); ACHR, Art. 9; ACHPR, Art. 6; Convention, Art. 7.

[274] *GIEM SRL and others* § 248-261.

[275] “No one shall be held guilty of any criminal offence on account of any act or omission which did not constitute a criminal offence under national or international law at the time when it was committed […]” (Convention, Article 7).

[276] See, among others, *Del Río Prada v. Spain* [GC], no. 42750/09, § 91, ECHR, 2013.

[277] G. v. France, no. 15312/89, § 25, ECHR, A325-B.

[278] Stark, (2016), 179-86.

inally liable" for such an act.[279] Put differently, persons should not be surprised to find that they are criminals.[280]

Now, for the Court, foreseeability must be appraised from the point of view of the defendant.[281] Whether a prohibition is foreseeable or not, depends not only on the instrument's content or the field it is called to cover, but also on the status of its addressees and the kind of persons it is targeting.[282]

Since foreseeability thus depends on its addressees, the Court has considered that it would be inconsistent to allocate blame without fairly attributing the offence to the blamed person.[283] Defending the contrary would leave that person in the position of someone who cannot avoid criminal blame. What does fair attribution mean? Here is where the presumption of innocence, enshrined in Article 6 (2), joins the scene.

Simply put, the Article demands to presuppose a person's innocence. Whether she is guilty, or not, hinges on demonstrating that such a supposition is wrong. The point of departure is always a person's innocence. Now, innocence would not be anymore the point of departure if someone is deemed liable despite not having any link with the case at stake.[284] Targeting with blame someone whose connection with the event cannot be established is equivalent to assume her guiltiness, thus twisting the presumption enshrined in Article 6 (2).

Consider, as an example, the Court's landmark case on the matter: GIEM SRL and others v. Italy. The case is a complex imbroglio involving several for-profit organizations and individuals. What is remarkable for the piece's purposes is that the Court found a violation of Articles 7 and 6 (2) because the companies were punished ─their goods were confiscated─ even though no judicial authority determined that they committed an offence. Criminal law "spoke"

[279] *Cantoni v. France*, no. 17862/91, § 29, ECHR, 1996-V. Cf. HrCtte., no. 2155/2012, *Paksas v Lithuania*, CCPR/C/110/D/2155/2012, 2014, where the Committee dismissed the claim of unforeseeability because the impeachment process did not lead to a criminal conviction. However, it found a violation of Article 25 (right to political participation) because the rule-making process lacked foreseeability and objectivity (§ 3.9, 7.8, 8.4).

[280] Stark Stark, (2016).

[281] *Kononov v. Latvia* [GC], no. 36376/04, § 235, ECHR, 2010.

[282] Ibid.

[283] *GIEM SRL and others* § 248-261. Remarkably, the Court grounds the point on the basis of a literal reading of Article 7 and the rationale of punishment. In the Court words. The 'penalty' and 'punishment' rationale and the 'guilty' concept (in the English version) and the corresponding notion of 'personne coupable' (in the French version) support an interpretation of Article 7 as requiring, in order to implement punishment, a finding of liability by the national courts enabling the offence to be attributed to and the penalty to be imposed on its perpetrator. Otherwise the punishment would be devoid of purpose (Varvara v. Italy, no. 17475/09, § 71, ECHR, 2013).

[284] *GIEM SRL and others* § 251.

with them without firstly determining that the act was theis. On the contrary, it assumed that its director's blameworthiness sufficed to also deem the organization blameworthy. And here was the problem. Even if the law is accessible, companies would not be able to avoid blame if a state willy-nilly deems them liable without establishing any connection with the event that led to a wrongdoing.[285]

### What sets the link between a person and her offence?

Now, the question is how to establish that kind of connection between a person and an offence? What is needed to defeat the presumption that someone is not blameworthy? The Court's initial wording is confusing. It argues that the linkage must be through a "mental link"[286]

It thus seems that the Court has in mind the same view of criminal liability that attribution gap theories presuppose. For the Court, as for the prevailing position, a fitting candidate would be one who caused an outcome with knowledge and control over its development. Hence, an attribution gap would be unavoidable if an unpredictable, autonomous or complex robot causes harm.

Undoubtedly, decisions link persons particularly clearly to their acts.[287] However, one can also make the case for punishment for not-being-aware and not-intending harm. Making that case requires that such not-being-aware and not-intending reveals relevant aspects of a person's character and dispositions. [288] Put differently, that there should be a mental link with the wrongdoing does not mean that the person must be fully aware and intend it.[289] Mental connections also exist whenever one's inner character, like carelessness or lack of control over one's passions, leads to an offence despite not being aware of it.

In further spelling its assessment, the Court seems to agree with that approach. Indeed, it speaks of mental linkage without specifying whether it *must* involve awareness and intent.[290] Furthermore, in spelling out the restrictions, it specifically argues that "in particular, the Contracting States may […] penalise a simple or objective fact as such, *irrespective of whether it results from criminal intent or from negligence*."[291] If one understands "negligence" as pre-

[285] Ibid., § 8-41.

[286] Ibid., § 242.

[287] Stark, (2016), 179-86.

[288] Ibid., 268-95; Hruschka, (1986).

[289] Something which would be impossible anyway. See Alexander, Ferzan, and Morse, (2009), 81-85.

[290] *GIEM SRL and others* § 251.

[291] Ibid., § 243.

cisely the failure to make the right choices, it seems that the Court is also considering it as capable of forming a mental link.

If establishing a mental linkage does not mean that the person must be fully aware and intent it, then what does it mean? How could one be sure that whatever a machine does reflect something about the inner character of the person? The Court is logically silent in this respect. Still, it is possible to depart from such a standard to build a set of baseline requirements. Before doing so, however, it is necessary to briefly address the restrictions to such a "human right to a personalized blame."[292]

### The scope of restrictions

The right to a mental linkage ─be it intention and awareness or carelessness─ is not unlimited. In *GIEM SRL and others*, the Court acknowledges that such a requirement "do not preclude the existence of certain forms of objective liability"[293] where a defendant's guiltiness is presumed without a court needing to demonstrate a mental link. As the Court put it, "the Convention does not prohibit such presumptions in principle; it does, however, require the Contracting States to remain within certain limits in this respect as regards criminal law."[294]

What are those limits? To what extent could a state get rid of the mental link in blaming someone? The Court has dealt with cases where a publishing director's liability was presumed for defamatory statements made in a radio programme in which he did not even participate;[295] or where a person's suffered a tax surcharge without her mental link with the offence being proven.[296] What the Court has demanded in these cases is to strike a balance between the importance of what is a stake ─the protection of other's honour or the effectiveness of the tax administration in the examples above─ and the right of a person to defend herself from the charges.[297]

Striking such a balance includes two demands. First, keeping no-mental link liability only to the extent that is strictly necessary to protect the interest at stake.[298] One could think, for instance, on petty offences that do not carry any stigma yet need to be punished swiftly.[299]

[292] Cf. Panebianco, (2014), 53-61.

[293] *GIEM SRL and others* § 243.

[294] Ibid.

[295] *Radio France and Others* v. France, no. 53984/00, § 24, ECHR, 2004-II.

[296] *Västberga Taxi Aktiebolag and Vulic* v. Sweden, no. 36985/97, § 113, ECHR, 2002.

[297] Janosevic v. Sweden, no. 34619/97, § 101, ECHR, 2002.

[298] Ibid.

[299] See Simester, (2005), 25-30. Against the possibility to render objective criminal liability as proportional, see Salako, (2006); Katz and Sandroni, (2018).

The second condition entails, as the Court clarified in *GIEM*, affording the individual the possibility to exonerate herself from the charges.[300] For instance, in *Salako* —where the applicant faced conviction for merely possessing prohibited goods when passing through customs—[301] the Court failed to find a violation of Article 6 (2) because the domestic court's assessment weighed his disregard to the warnings issued before taking the goods with him, while the law would allow acquitting him if he succeeded in demonstrating force majeure.[302]

Ignoring any of these limitations would indeed entail absolutely depriving an agent of her right under Article 6 (2), thus overstepping the Court's permission to limit defendant's rights.[303] It is tantamount to punishing —and stigmatising— without getting any personal.

Those restrictions will come back as the thesis models the scenarios of blurred accountability. For now, it is important to understand that, even if presuming a defendant's liability is allowed, such a presumption can only occur within strict limits. If a state decides to secure its duties to redress by blaming a developer or robot user despite her participation, it should be sure that doing so is (i) necessary to secure the fulfilment of its obligations to criminal redress and (ii) leaves room for the defendant to escape blame. Failing to secure both aspects would render any mobilisation of criminal laws contrary to human rights standards, at least as depicted within the ECtHR.

### *ii. Getting personal with the human "behind" the machine*

How could one go beyond the machine and target a person who did not know nor intended what an unpredictable and autonomous machine would do? How could one establish a "mental link" in the complex imbroglio that those artefacts imply? The thesis submits that three baseline conditions are necessary. If a court demonstrates that the unaware developer or user meets them, it would be right in blaming her for whatever the artefact does.

#### A failure to form a belief upon the risks that a robot was creating

The first of those conditions is a *failure to form a belief upon the risks that a robot was creating*.[304] Here, one need to combine two elements. First, one need to consider everything the person knew before developing or using the robot. Here is where technical standards and pro-

[300] *GIEM SRL and others* § 243.

[301] *Salabiaku v. France*, no. 10519/83, § 26, ECHR, 1988 A141-A.

[302] Ibid. § 30

[303] Ibid. § 28.

[304] Stark, (2016), 229-43.

fessional codes of conduct feature.[305] However, that is not enough: one need to go further and grasp everything a person knew before engaging with the machine. If a police officer uses a robot to control a protest and things go wrong, one should ask: did her codes of conduct set some boundaries on the use of tools to control demonstrations? Did her training provide information on robot's unpredictable behaviour within crowded environments?

The second component is crucial, however. One is not blamed for her own lack of background knowledge, but for failing to put things together when things went wrong. Thus, the second requirement demands a perception of the circumstances at stake. To blame the officer in the example above, a court needs to ask a further question: assuming her training allowed her to know that robots might exhibit erratic behaviour when left in crowded environments, did she perceive that the demonstration had enough people to be concerned about the device going wrong? Was she somewhat impaired – imagine a situation of emergency – to do so?

Those two questions will render the first element. If a court can establish that (i) an agent had enough background knowledge and (ii) her perception of the events were such that she could put things together and determine that things would go wrong, it could say that she *failed* to form a belief upon the risks that her actions – here, deploying a robot in a demonstration— were creating.

### A failure of belief that says something about the agent's character

However, a person can fail to form a belief for reasons that are strange to her character. Findlay Stark thus suggests to further ask whether the feature of the defendant's motivational set-up that prevented her from forming a belief is in fact reflective of her as an agent.[306]

Such a "motivational framework" refers to an agent's overall character, including not only her specific dispositions when things went wrong, but her general character.[307] One need to know whether such a failure reveals a general disposition to show insufficient concern for other's interests.[308] In this vein, character traits become the focus insofar as they reveal lack of sufficient care for the legally protected[309] interests of others.

[305] See Horder, (1997). He proposes focusing on the exclusionary reasons flowing from the guidelines that regulate an activity one voluntarily undertake. Cf. Hart, (2008), 136-57.

[306] Stark, (2016), 247-52.

[307] Ibid., 248-50.

[308] Ibid., 247.

[309] Certainly, one cannot affirm a failure to take care o finterests that bear no legal protection. See ibid., 261-66.

What counts as “sufficient” concern depends on the standards of each society and on the interest at stake.[310] However, one cannot expect a person to always investigate all potential risks. In those circumstances, it would be difficult for citizens ─however well-intentioned─ to avoid being culpable. This would be tantamount to blurring the presumption of innocence. Indeed, a too stringent notion of sufficient concern would put citizens in the position of one who cannot exonerate herself. If to show sufficient concern a developer of LAWS must consider each risk of her development, she would inevitably fall short of the standard and would end up in a position where she cannot dispel the thought of her failure.[311] If showing sufficient concern is too difficult, everyone would automatically fall short of it.[312]

Probatory hurdles aside, the main point here is that a court should ask whether a person was careless in developing dangerous tools, or whether she tended to show insufficient concern in deploying those machines. In general, the question is if there is match between the behaviour that ended up with a machine harming and the offender’s character insofar as the latter speaks of her concern for others. Put differently, such a failure should not be alien to the person “behind”; it should not spring from somewhere outside the framework of dispositions towards thought and behaviour that reflect her care for others.[313]

Going back to the example above, one need to add a further question once it is established that the officer failed to form a belief upon the risks of unpredictable behaviour stemming from using a robot in a protest. That question zooms into her persona, and asks, whether that failure says something of her. Is she the kind of person who would exhibit the kind of insufficient concern that ended up with a robot causing havoc in a demonstration? Or is it a feature that is rather alien to her character?

Doing so ensures that the connection between a defendant and her wrongdoing is sufficient personal. It is the “mental” element of the link that the ECtHR demands. If one only relies on the first condition, criminal liability loses its focus on the defendant. It instead turns into blame for failing to meet an expectation –forming a belief of the circumstances— but says

[310] Third, the requirement is not fully objective nor fully subjective. Whereas the former would fail to reach the defendant’s persona, the latter would made prohibitions dependent on the defendant’s own appreciation. Hence, whereas concern is assessed in light of a person’s own traits, sufficiency is determined in light of an objective standard (ibid., 257-59.).

[311] Some scholars stop here and assume that the key to attributing liability is the definition of the level of sufficient concern when using machines. Dealing with it as a case of ‘normal risk of daily live,’ Gless et al., (2016), 432-33.

[312] Alexander, Ferzan, and Morse, (2009), 81-85.

[313] Stark, (2016), 251.

nothing about her subjectivity. [314] Hence, the need to add this requirement as a baseline element for blaming the unaware person behind the machine.

### Things would have been different had she displayed sufficient concern

However, it is not enough to say that a defendant had such and such character traits; it is also necessary to connect them to the specific wrongdoing.[315] That renders a further baseline condition: one must ask what would have happened if the defendant had not had the traits that led her to failing to form a belief. Specifically, had the defendant showed the necessary traits leading to forming a belief upon the circumstances at stake, would she be able to impede the materialisation of harm.[316]

The requirement is an important one. It points to the risk that focusing exclusively on an agent's traits might create a presumption that is unacceptable under human rights standards spelled out above.[317] Indeed, it would lead to presuming that, because an agent is so and so, she *must have* given rise to the harm. Put otherwise, it turns criminal law into an appraisal of personal features. And, insofar as it is divorced from the concrete facts, how could she show that she is not as the court framed her? Those difficulties make it hard to imagine such a restriction capable of acceptable from a human rights perspective. [318]

The point also serves to counter two potential objections. Indeed, some scholars argue that focusing on character traits is susceptible to the "significance in action problem." According to it, an agent is appraised only for his personal features, whereas their relation to the ensuing harm become irrelevant.[319] By defending the need to ask what would have happened if the defendant had shown more concern, that problem is avoided. An agent that develops a robot is not blamed for being careless, but for the fact that her carelessness is the reason why harm ensued.

The other objection has to with the fallacy of division. It occurs when one reasons that something that is true for a whole must also be true of all or some of its parts. [320] Here, it would entail saying that, because a robot —the "whole"— is dangerous, all of the developers who intervened —the "parts"— also are. Demanding a link with the facts answers this objection. An agent does not receive blame just because she is part of the complex network behind the

[314] Ibid., 181-86.

[315] Alexander, Ferzan, and Morse, (2009), 72-73.

[316] Hruschka, (1986), 691-700.

[317] See Sub-section i.

[318] Bandes, (2010), 447.

[319] Stark, (2016), 253.

[320] Velasquez, (2003), 540-41.

harmful robot. Nor is she liable just because failed to display concern *while* being part of that network. There need to be a connection between her behaviour and the robot's harm. Insofar as robot's behaviour is the consequence of unforeseen interactions between its parts and being more diligent would have not impeded that behaviour, no offence should be attributed to the nodes.[321]

Hence, it should be almost certain[322] that showing sufficient concern while forming a belief upon the ensuing risks would have put the defendant in the position necessary to avoid their materialisation. If there is a doubt, the defendant should be spared. Otherwise, again, the defendant would be presumed to have given rise to a wrongdoing just because she had certain features. That is the last condition to blame a not-knowing defendant behind a machine.

### *iii. Improving the first step of the attribution gap concept*

Section A above defended, first, the need to focus on the person behind the robot and, second, to broaden the vision of criminal law underpinning prevailing positions. Now it is necessary to take the conditions introduced in Section B to generate such a wider vision.

Recall that the first condition for the mismatch that makes up an attribution gap is as follows:

> (1) Under their vision of criminal law, someone is accountable provided that she (*i.*) caused a particular outcome, (*ii.*) with knowledge and control over its development, and (*iii.*) with the ability to understand the implications of her behaviour.

In building upon ECtHR's demand of a mental link, one could improve it it in the following manner:

> (1) someone is accountable provided that she
>
> a. (*i.*) caused a particular outcome, (*ii.*) with knowledge and control over its development, and (*iii.*) with the ability to understand the implications of her behaviour; or
>
> b. (i) failed to form a belief upon the risks of deploying or developing a robot; (ii) such a failure being reflective of her general display of insufficient concern for other's interests, and (iii) it is almost certain that she would have avoided such a failure had she shown more concern in forming her beliefs.

[321] Here, the assessment is not one of causality. On the contrary, it is a counterfactual hypothesis: what would have happened *if* the agent had shown more concern. See Reyes Romero, (2015), 150.

[322] Otherwise, again, the state risks presumming an agent's culpability. See ibid.

Attribution gap theories would further focus on scenarios where (2.1) either there is no fitting candidate to respond, or (2.2.) no one responds to the degree that it would have if no robot were present. As the next Section shows, that second part of the mismatch also needs being reframed. A combination of both changes would render a model that yields a more accurate picture of the fate of criminal accountability in the age of robots and AI.

## C. A model of attribution failures: towards a model of technologically blurred accountability

Ask when do robots present an attribution gap, and the reply will be that it is when there is no one to respond. Indeed, the prevailing position presents only two scenarios: either no one is accountable, or no one is accountable to the expected degree. Sub-section i. argues that such an approach is flawed. On the one hand, because it departs from an erroneous reasoning. On the other, because it fails to capture those circumstances where a lack of accountability frustrates human right standards.

Sub-section ii. develops an alternative account. It does so by developing the idea of "fitting candidate" in the definition of attribution gap. For attribution gap theories, no fitting candidate equals no candidate at all. Sub-section ii. instead proposes three scenarios where deeming someone liable would frustrate obligations to redress—either because no one is liable or because deeming someone liable would unjustifiable sever the mental link that the ECtHR has demanded. A *fitting* candidate is one who can be called to account for a lack of concern for others as expressed in her failing to see that using or developing a robot would render harm. Deeming liable someone who falls short of that standard without any justification blurs accountability as much as not blaming anyone.

### *i. From attribution gaps to a model of technologically blurred attribution*

Attribution gap theories seek to show how the introduction of AI and robotics blocks the path to attribution. In so doing, their main argument is that *no fitting candidate* for liability is to be found.[323] However, their reasoning is flawed and says nothing about the circumstances where a lack of accountability frustrates human right standards.

It is convenient to start with the flaws. The prevailing position does not say who can be a fitting candidate for accountability. Instead, it follows a negative argumentation. Typically, it starts with a list of potential candidates ─normally developers, manufacturers and drivers (the latter, in scenarios involving automated vehicles). Then it contrasts different ways of attributing liability to show that such an endeavour is doomed to fail. Common bones of contention

[323] See Chapter 3.

are models of direct responsibility, indirect responsibility or command responsibility. And they go through the requirements of each of these models to show that neither the driver, nor developers, manufacturer and so on, are fitting candidates for responsibility.[324]

The problem with such a method is that it is always possible to find a different ─perhaps better─ model for attribution or an alternative candidate to respond for the criminal offence. That a specific model of attribution ─command responsibility, for instance─ fails to deliver or that the commander cannot respond, says little about the existence of a gap. Indeed, it does not follow from the fact that a developer or users are not liable that there is no fitting candidate out there to deem accountable.

Crucially, however, the model is ill-suited for spotting the circumstances where human right standards end up impaired. This is because the prevailing position only sees a gaps when no one is deemed accountable –be it that the person is completely spared or only targeted with a reduced punishment. However, that appraisal fails to capture what is essential of the "coercive sting" of human rights law. Showing that benefits from using the following case:

> *An agent A uses an AI-based interrogation system at a border. After spotting suspicious behaviour, she invites a person to go into a room where a device will ask some questions. The technology has a voice recognition system that memorises every utterance aspect of the interaction. It is programmed to spot inconsistencies and mention them while interacting with the interrogatee. In using some of those patterns with the interrogatee at stake, it roused such a feeling of anguish and inferiority that he ended up declaring a wrongdoing he did not commit. Later investigations also demonstrated psychological damages.*[325]

Now imagine the officer skips criminal blame –or she gets a lower punishment. The prevailing position will stop here and call that situation an "attribution gap." However, that lack of punishment might be because harm ensued despite she employed all the safeguards to avoid causing suffering. Or a court might accept the defence that she acted under duress. Indeed, domestic systems include different circumstances that block punishment. Think of honest errors or self-defence, or situations where a prosecutor fails to demonstrate the defendant's involvement. If verified, these circumstances impede punishment. And there is no breach of human right obligation.

The problem with the prevailing position is that it provides no criteria to distinguish those cases where punishment is withheld from those where a state fails to comply with its obliga-

[324] See, e.g., HRW and IHRC, (2014).

[325] See McAllister, (2017), 2540-45.

tions of redress. Sparing the agent because a court acknowledged her diligence does not entail a failure to meet obligations to redress. That would occur only if the framework is such that it falls short of its deterrent function. However, deterrence does not demand punishment at all costs, but only when it is fair to impose it on a blameworthy candidate.[326] That is precisely the distinction that prevailing positions miss.[327]

Again, what if someone, whoever it is, is deemed accountable. Imagine the state deems the officer liable despite her display of care. Or that an inadvertent programmer ends up targeted. Could one assume that there is no gap at all? Certainly no. That individuals have a right to have a criminal law framework is not equivalent to have any kind of blackletter law. As built within the ECtHR, the framework needs to be capable of deterring undesirable behaviour, something that presupposes a fair distribution of blameworthiness.[328] Hence, targeting someone despite her innocence is as problematic as not targeting anyone. Once again, the prevailing position's idea of a gap fails to distinguish between cases where human rights end up frustrated and those where the state rightfully withholds punishment.

Moreover, if followed to its ultimate consequences, the prevailing position risks turning "the coercive sting of human rights" into "coercive overreach."[329] Implicit in their notion of a gap is the idea that some fitting candidate should be out there whenever a robot causes harm. That definition conflates the conditions for criminal responsibility with those needed to ascertain a human rights violation. In the example above, that would mean that for each case of ill-treatment –a violation of Convention's Article 3– someone needs to respond. Such an approach overlooks the different structure and consequences of these areas of law, and thus neglects the special principles necessary for blame and punishment of individuals.[330]

### *ii. Improving the second step of the attribution gap concept*

How to provide an account that avoids the flaws of the prevailing position, while being broad enough to cover the situations where duties of redress end up frustrated? This Sub-section develops the notion of *fitting candidate* (2) and expands the scenarios (2.1/2.2.) to those cases where machines turn accountability into an exercise falling short of human right standards

[326] See Chapter 4.

[327] Particularly, HRW, wich argues that "In both law enforcement operations and armed conflict, the actions of fully autonomous weapons would likely fall within an accountability gap that would contravene the right to a remedy. *It is unclear who would be liable when an autonomous machine makes life-and-death determinations about the use of force without meaningful human intervention*." (emphasis added) (HRW and IHRC, (2014), 19.)

[328] See Chapter 3.

[329] Lazarus, (2012), 136, 47.

[330] Robinson calls this "substantive and structural conflation" (See Robinson, (2008), 925, 29.) See also Stoyanova, (2014).

(2.3.). In so doing, it makes the definition of attribution gaps reflective of those circumstances where technology frustrates duties of redress.

Who could be a fitting candidate? Which are those scenarios where the criminal apparatus misses her? These are two questions are central to improve the second step of the attribution gap concept. Indeed, in many circumstances where a robot causes harm, the problem will be not that no one *is* called to account, but that those who are *should not*.

Following the ECtHR's jurisprudence, one could argue that a fitting candidate is one whose mental connection with the offence is established.[331] Whether it is direct liability, or command responsibility, is a secondary matter. A fitting candidate if one with whom the criminal law can get personal.

That answer already points to those scenarios where attribution gaps would arise. The mismatch is not so much between a model of attribution and a failure to meet its requirements, as prevailing positions assert. On the contrary, it is between the human rights-based duty to set a criminal law framework that targets the defendant, and a failure to establish the mental link needed to do so. In which scenarios would that happen?

The model submits that instances of blurred attribution will occur insofar as imputing criminal liability would lead to (i) rendering criminal prohibitions unforeseeable; (ii) severing the mental link between an offender and the wrongdoing, or (iii) frustrating the human rights-based obligation to set forth a legal framework. Overall, a candidate is unfit to respond insofar as it is illegitimate, on human rights ground, to deem her liable (or hold a conviction) for the behaviour of a robotic system.

The first type of blurred accountability focuses on the foreseeability of prohibitions. Recall that the need for a mental link supposes an obligation to set foreseeable prohibitions. One should be able to avoid criminal liability. Hence, an unfitting candidate is one who is blamed for prohibitions that she could not have known. When would that happen when a robot intervenes in an offence? Chapter 6 argues that, at least, that would be the case when developers are called to verify the autonomous behaviour of machines expected to engage in the most unfamiliar activities.

The second scenario points to cases where criminal liability cannot get personal. The reason is that robots impede establishing a mental link with the offender. These scenarios cover cases where a court cannot say that whatever the robot did is reflective of an individual's lack of concern. Again, Chapter 6 introduces one case where this scenario could materialise.

[331] See Section B (i.-ii.).

The cases clustered within the third category present an scenario where the state refrains from blaming someone for the machine behaviour. To the extent that criminal liability fails to target the person who is linked to the offence it would fail to set forth a framework that meets the standards of proportionality, dissuasiveness and effectiveness.

Having the three scenarios, one could improve the second step of the definition where attribution gaps arise in the following manner:

> (2) Whenever a robot harms it is impossible to find a suitable candidate because the latter's mental link with the offence cannot be established, be that
>
> > either because (*i.*) the machine leads to a complex setting; (*ii.*) or because it is unpredictable; (*iii.*) or because it robot itself fulfils the conditions for being a suitable candidate; or (iv.) because of any other reason impeding the determination of  a mental connection between the agent and the robot's wrongdoing.

Such a situation leads to a scenario where (2.1.) those who are called to account lacked the possibility to foresee the prohibition they supposedly infringed; or (2.2.) the wrongdoing is not a reflection of their lack of concern or (2.3.) the  legal frame-work would fail to attribute liability to the fitting candidate despite being the case that the offence is generally punished or led to situations of intentional/grossly care-less life-deprivation, severe injuries and attacks against fundamental aspects of an individual's privacy.

To set it apart from the definition of attribution gap, one could call this model as one of "technologically blurred accountability."

## D. Conclusion

This Chapter introduced an alternative model –technologically blurred accountability— to characterise instances where robots would cripple accountability. It started by criticising the first step of the definition: that pointing to the prevailing vision of criminal liability. In this sense, Section A suggested shifting attention to the person behind, who can be liable despite being unaware of the harms that the robot causes.

Section B explained how one could do it. In spelling the implications of the obligation to establish a mental link, as presented in the ECtHR's jurisprudence, it set the stage for Section C. That Section focused on the second step of the definition of attribution gaps. After contesting the focus of the prevailing position, it proposed three scenarios where a robot might impede finding a fitting candidate for liability.

The takeaway of the Chapter is thus an improved definition of attribution failures —the model of technologically blurred accountability. The improved definition goes as follow:

*(1) someone is accountable provided that she*

- *a. (i.) caused a particular outcome, (ii.) with knowledge and control over its development, and (iii.) with the ability to understand the implications of her behaviour; or*

- *b. (i) failed to form a belief upon the risks of deploying or developing a robot; (ii) such a failure being reflective of her general display of insufficient concern for other's interests, and (iii) it is almost certain that she would have avoided such a failure had she shown more concern in forming her beliefs.*

*(2) Whenever a robot harms it is impossible to find a suitable candidate because the latter's mental link with the offence cannot be established, be that*

*either because (i.) the machine leads to a complex setting; (ii.) or because it is unpredictable; (iii.) or because it robot itself fulfils the conditions for being a suitable candidate; or (iv.) because of any other reason impeding the determination of a mental connection between the agent and the robot's wrongdoing.*

*Such a situation leads to a scenario where (2.1.) those who are called to account lacked the possibility to foresee the prohibition they supposedly infringed; or (2.2.) the wrongdoing is not a reflection of their lack of concern or (2.3.) the legal framework would fail to attribute liability to the fitting candidate despite being the case that the offence is generally punished or led to situations of intentional/grossly careless life-deprivation, severe injuries and attacks against fundamental aspects of an individual's privacy.*

The next two Chapters set that model in action. In Chapter 5, it shows that robot's unpredictability, complexity or autonomy do not block criminal accountability. In turn, Chapter 6 takes it to identify two cases where AI and CPSs will introduce scenarios of technologically blurred accountability.

# 5 The model in action: disentangling the puzzle of attribution gaps

This Chapter sets the model developed above in action. It does so to show that most of the cases presented in the literature as posing an attribution gap, are not as problematic as they appear. In fact, as the Chapter argues, it is possible to attribute liability without unduly restricting defendants' rights nor frustrating human rights duties of redress.

The Chapter depicts the argument by presenting five cases. Section A tests cases of unpredictability. Section B, in turn, focuses on cases of autonomy. Lastly, Section C tests the model on a case of complexity. The three Sections follow a similar structure. After introducing the case, the model of technologically blurred accountability is used to determine whether robots exhibiting these features introduce one of the three scenarios described in the previous Chapter. A different Sub-section explains why the explanation behind each of the variants is flawed.

## A. Unpredictability and blurred attribution

Much of the existing research points to foreseeability as the greatest challenge that AI and robotics pose for criminal law. That robots are "unpredictable by design" serves to ground the argument that these devices block attribution.[332] Contesting that approach demands appraising two cases of AI's unpredictable behaviour under the model pf technologically blurred accountability (Sub-section i,). After that appraisal, Sub-section ii. further asks whether the kind of "unpredictability" that these technologies present would exclude criminal liability.

### *i. Unpredictability as a source of technologically blurred attribution*

The robotic butler

Consider the following fictitious case:

> *A company selling artificial intelligence software sells its product to a racist. The racist proceeds to install the software onto a robot butler, and the robot butler proceeds to learn and develop under the teachings of its owner. One day, a black UPS driver delivers a package to the front door. The now-racist robot answers the door and, upon seeing the black UPS driver, thinks, "the only reason a black person would be on my front porch would be if he were here to burgle my owner." The robot proceeds to attack the UPS driver under the mistaken assumption that he is a burglar.*[333]

[332] See Chapter 2. Calo, (2015), 543. See also Millar and Kerr, (2016).

[333] Kowert, (2017).

This situation presents a case of reinforcement learning, that is, the machine learns through its interaction with the environment. The only thing a designer embeds is a goal ─here, being a good butler.[334] The kind of input it takes, and whether is correct or not, is something out of the control of the human "behind." A typical case of indeterminate and arguably unforeseeable behaviour.

Is the user criminally liable for the robot's attack? The prevailing position would give an affirmative answer. They would focus on the first part of the definition and say: criminal attribution requires that the agent caused an outcome with knowledge and control over its development (1.a.). One cannot know nor control what one cannot foresee. Machines that learn by itself are unpredictable, and thus their behaviour cannot be known nor controlled. Hence, whenever a robot harms it is impossible to find a fitting candidate (2.ii.).

Once one adds the model developed here to the definition, the answer is different. The question is not whether the agent knew or controlled, but whether it is possible to blame her for not knowing and not controlling. Take the user. The first question is: (1.b.i.) did she fail to form a belief upon the risk that such a self-learning device would end up attacking a mail carrier? And here are reasons for an affirmative answer. First, it is possible to ascribe her the background belief that devices learn upon interaction. Otherwise, it would present no value as butler in an unstructured environment. Furthermore, she might also know that CPSs can exert themselves upon the world, for there is no point in having a virtual butler. She needed –and thus knew— that CPSs can act in the physical world. Hence, knowing that the machine would absorb and change upon external input, while being able to exert itself in the world, seems to be part of the agent's toolkit.

Now, were the circumstances given so that she could actualise her background knowledge and form a belief upon the risk of harming a visitor (1.b.i.). Again, the answer could be affirmative. Upon perceiving the now-racist robot and its physicality, one could expect that, if left to receive visitors, it might cause harm. In noticing that she did not grasp those risks, a court can plausibly conclude that she failed to form a belief upon the risks her robotic butler was creating.

The next two questions can also receive a positive answer. Did such a failure to belief spring from a character trait demonstrating lack of sufficient concern? (1.b.ii.) One might be tempted to point to racism as the key feature. However, the assessment is not about who she is as a

[334] Reinforcement learning presents the problem of unpredictability in its strongest sense. Embedded with broad goals, the agent uses trial an error to come up with a better solution. It thus gets either rewards or penalties depending on how it maximises the goal at stake. (See Sutton and Barto, (1998).

person. What is relevant of her personality is whether some features demonstrate insufficient concern for the interests of others. Therefore, the key factor here is the carelessness in delegating the responsibility to receive visitors to a machine. Despite its learning abilities, such a robot seems uncapable of grasping social norms about the justification of using force. Hence, leaving such a self-learning robot with the physical power to cause harm roaming in a house is arguably a display of lack of concern for the interests of others.

Concerning the last element of the definition of criminal liability (1.b.iii.), the question is whether showing sufficient concern would have rendered a different outcome. And that seems to be the case here. Had she realised the risk of harm, she might have kept the robot within defined parameters, thus being able to prevent it from harming the mail carrier.[335] Indeed, one can keep unpredictable devices safe by restricting the scenarios where they can act. And doing so would have allowed the user to avoid injuring others.

The poisonous vehicle

Ryan Calo has offered the following hypothetical case:

> *Imagine one manufacturer stands out in this driverless future. Not only does its vehicle free occupants from the need to drive while maintaining a sterling safety record, it adaptively reduces its environmental impact. The designers of this hybrid vehicle provide it with an objective function of greater fuel efficiency and the leeway to experiment with system operations, consistent with the rules of the road and passenger expectations. A month or so after deployment, one vehicle determines it performs more efficiently overall if it begins the day with a fully charged battery. Accordingly, the car decides to run the gas engine overnight in the garage—killing everyone in the household.*
>
> *Imagine the designers wind up in court and deny they had any idea this would happen. They understood a driverless car could get into an accident. They understood it might run out of gas and strand the passenger. But they did not in their wildest nightmares imagine it would kill people through carbon monoxide poisoning.*[336]

[335] The ISO standard —arguably applicable to the robot butler in the example— include requirements to restrict the device's range of movement. As a risk reduction measure, it might be required either to constrain —via software or through other means— the space where the robot acts. The manufacturer shall state those capabilities in the instructions for use. One could then further ask whether, in the example above, the developers of the software failed to embed those restrictions and hence showed insufficient concern (see Robots and robotic devices — Safety requirements for personal care robots, ISO 13482:2014 (ISO, first published February 2014), 6.4). Cf. Bernhard et al., (2021).

[336] Calo, (2018), 34-35.

Does the developers' argument hold? Does the unpredictability of a robot capable of poisoning the household with monoxide negate criminal liability? They seem to be pointing to the conditions described in (1.a.ii.), so the prevailing position would have to accept their argument:  they lacked knowledge and control over the robot's behaviour. Hence, their behaviour attracts no liability.

However, the model introduced here suggests a different answer. Indeed, if one looks at the requirements introduced in 1 (b.), developers' ignorance is not a excuse but the main reason for their liability. As designers, indeed, it is plausible to ascribe them the belief that giving a car such a leeway to make decisions might end up with behaviour they could not expect (1.b.i.). The BS 8611:2016 standard already includes, as an ethical hazard, the risk that robots might develop new or amended action plans with unforeseen consequences.[337] Emergent behaviour was the price they paid to have a car which can adapt to the "rules of the road and passenger expectations," and industry codes already provide enough information on such a price. Hence, they could plausibly receive blame for failing to understand that, in some circumstances, such adaptability would end up in a wrongdoing (b.i). Once you give a car such a leeway, it seems plausible that killing a household with monoxide is something you should have in your calculus.

There are also strong arguments in favour of the two other elements of the model. Indeed, there is nothing in the technology that would invite to think that a court cannot make the assessment that the failure to restrict the car's option did not stem from the designers' character (1.b.ii.). Nor are there reasons to say that things would have not been different by showing more concern (1.b.iii.). On the contrary, they might have ended up with the alternative to impede the death of the household. For instance, the BS 8611:2016 would have recommended to design the car such that it can inform them when new forms of behaviour have been developed.[338] They could have also shown sufficient concern by testing the device or embedding additional security redundancies.[339]

Blurred attribution?

Now, would these two extreme cases present a situation of blurred attribution? The prevailing position would say that no one responds. However, as argued previously, that no one responds is not sufficient to assert a problem of attribution. On the contrary, the latter becomes problematic when it leads to a scenario where (2.1.) those who are called to account lacked the

[337] Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems, BS 8611:2016 (BSI, first published April 2016, currently under review): 4.4

[338] Ibid.

[339] Ibid, 7.

possibility to foresee the prohibition they supposedly infringed; or (2.2.) the wrongdoing is not a reflection of their lack of concern or (2.3.) the  legal frame-work would fail to attribute liability despite being the case that the offence is generally punished or led to situations of intentional/grossly careless life-deprivation, severe injuries and attacks against fundamental aspects of an individual's privacy.

Neither of these scenarios is realised in the cases above. Assuming the butler's user and the car developers are called to account, they cannot argued that criminal prohibitions where unforeseeable. The connection between them and their background beliefs makes it predictable that they were under a prohibition to, on the one hand, leave a potentially harmful robot alone and, on the other, design a machine which such a leeway that it can poison the household.

In what concerns the mental linkage, there are also reasons to trace back both failures to the defendants' lack of concern. Neither the fact that a robot mediated between their behaviour and the ensuing harm, nor that it was unpredictable, impedes to look at their display of insufficient concern. On the contrary, it is precisely the fact that such robots were used or developed what demonstrates their lack of concern for other interests.

Could the state argue that its legal system is ill-suited to couch these two cases of life-threatening injuries? Could it hold punishment on the basis that technology makes it difficult to attribute liability? Again, that is hardly the case: a legal system which allows blaming the inadvertent, careless subject, would be also well-suited to apportion liability when machines are involved.

Undoubtedly, cases with this structure will raise questions of whether there was enough evidence to attribute harm. At their best, however, these considerations highlight epistemic and practical problems common in criminal courts. They might certainly suggest the development of forensics that can capture instances of harm with robotic systems involved. However, they do not suffice to threaten duties of redress nor provide reasons to doubt that standard resources for the assignation of responsibility come up short here.

### *ii. Unpredictability and blurred attribution*

#### What does foreseeability mean?

Few would question the relevance of the idea that one should only answer for foreseeable harms. Undoubtedly, it is hardly surprising to see it underpinning the problem of attribution.[340] Certainly, the model does not mean that one can be liable for any event, no matter how unpredictable it is. Yet, the kind of (un)foreseeability that negates the attribution of criminal liability is different from the kind that robotics and AI pose.

In both Common Law and Continental Law systems, harm is foreseeable to the extent that is the kind of harm one has in mind when defining activity as risky. Foreseeability is a generalisation of specific activities and the ensuing risks. Thus, each risk involves a "model hazard,"[341] and the question is whether that "model hazard" is the same kind of hazard as the "actual hazard" materialised in the case at stake. Return to the examples above. In investigating whether the injuries were foreseeable or not, one needs to ask whether harming the carrier — an actual hazard— is the kind of harm one would typically expect from leaving a now-racist machine with the responsibility to attend visitors —model hazard.

That account has two implications. First, a defendant need not foresee every single detail of the outcome.[342] Whoever devised a car which can optimise its fuel consumption need not to foresee that it will kill such a household in such circumstances. The second implication is that foreseeability is an objective standard. It is irrelevant whether the agent at stake was actually aware of the risk or not.[343] The question is whether the ensuing injuries match the type of hazards any reasonable person would expect from developing a device with such a leeway.

Are robots unforeseeable?

To the question "are robots predictable," authors like Ugo Pagallo,[344] or HRW, [345] seem to answer, "no, they are not." And they are right: these devices are developed to gather an increasing amount of data, and there is no model to predict outputs from all the possible inputs. Furthermore, they learn how to handle the input through several iterations, such that the rationality behind decisions cannot be explained.[346]

One can draw on the theory of "cognitive uncontainability" to explain this feature. Simply put, an artefact is cognitively uncontainable because it is smarter than the human behind it.

[340] Among others, the CoE is one of the institutional voices pointing to that concern. See Presidency Conclusions - the Charter of Fundamental Rights in the Context of Artificial Intelligence and Digital Change, (2020). para. 5

[341] Hart and Honoré, (1985), 254-90, 432-64. In the realm of robotics and AI: Selbst, (2020), 1342-45; cf. Gless, Silverman, and Weigend, (2016), 427.

[342] Hart and Honoré, (1985), 255-190; Beck, (2016), 139.

[343] Hart and Honoré, (1985), 255-190; See alsovan der Wilt, (2015); Vanacore, (2015).

[344] Pagallo, (2013).

[345] HRW and IHRC, (2014); Human Rights Watch & International Human Rights Clinic of the Human Rights Program at Harvard Law School, (2012); HRW and IHRC, (2015); Human Rights Watch, (2020).

[346] Millar and Kerr, (2016).; Burrell 2016: 1– 12; Rahwan et al., (2019).

That might be the case because it knows facts a person cannot know —known as “strong uncontainability”— or because the device is operating in a rich domain and searches a different part of the space that humans find difficult to search.[347] In this sense, a rescue robot that inspect a hazardous area inaccessible to a human would be uncontainable in that it will grasp and react to facts that remain unknown. In turn, IBM’s Deep Blue, which in 1997 defeated the then-world chess champion, uses strategies beyond its opponent’s grasp. Otherwise, it would have been possible for him to defeat it.[348]

The theory of cognitive uncontainability shows why the problem of machine unpredictability is misplaced. The problem is not whether robots are unpredictable or not. Criminal law is instead interested in knowing whether the harm caused is the kind of consequences one would expect from robots’ features and the context they are used. IBM’s Deep Blue was certainly inscrutable. No one knew which kinds of movements it would come out with. And yet, nobody would have expected it to issue harassing messages to its opponent as a tactic or leaving the table, for instance.[349] Such behaviour was not within the kind of consequences that the activity —playing chess— entails.[350]

Compare to the case of animals intervening with independence of humans. Like robots, whether such an animal excludes responsibility depends on whether the intervention is so abnormal to exclude responsibility.[351] Animals also are unpredictable to an extent. One might not know whether a barking dog is to attack, nor whether one would encounter a violent moose in the forest. However, the extent to which that behaviour excludes attribution does not depend on that kind of foreseeability. On the contrary, the question is: does it goes beyond what one could expect of “the specific nature of that animal.”[352] If the answer is no, then the behaviour is within the model hazard and is thus foreseeable.[353]

To the extent that machine’s unpredictability is like animals’, it does not exclude the attribution of criminal liability. The pertinent question is not whether robots are more or less predictable but whether their use in a specific context poses risks of a particular kind. If risks of that type materialise, it is irrelevant whether the behaviour was emergent. Attribution will be excluded only when the upshot goes beyond the “model risk” that such a use entailed. How-

[347] Yampolskiy, (2019), 110.

[348] Ibid., 114.

[349] See also Selbst, (2020), 1343.

[350] Hsu, (1999).

[351] Hart and Honoré, (1985), 347-49.

[352] Ibid.

[353] Also Selbst, (2020), 1344-46.

ever, nothing in the development of technology nor in scholars' examples suggest that that is the case.[354]

### B. Autonomy and blurred attribution

Now, it is in order to see the model dealing with cases of autonomy. Going through all the examples would certainly go beyond this piece's purposes. Here again, Sub-section i. discusses two prototypical cases of robots exhibiting autonomous behaviour. Sub-section ii. shows why focusing on autonomy provides an erroneous account of criminal accountability.

#### *i. Autonomy as a case of blurred attribution*

The case of LAWS

Killer robots have been the main, if not the sole concern of human rights organisations. Particularly known are HRW's reports justifying their ban in light of, among other things, the gap in attribution of liability.[355] Recently, moreover, a UN-SC report made it to the headlines when it narrated how autonomous drones presumably engaged some targets during the Second Libyan Civil War.[356] One can thus speculate, as HRW does,[357] whether similar situations will replicate within the context of law enforcement activities. If so, as HRW argues, one of the hurdles is that there will be no fitting candidate for responsibility.[358]

Imagine the following scenario:

> *A device is equipped with sensors to perceive the environment, a data base of trillions of cases where lethal force has been necessary and a learning algorithm that helps it to predict whether an individual is posing an immediate risk of physical injury or death to others. If the device classifies an individual as posing such a threat, it has the necessary capabilities to use lethal force without requesting authorisation. An agent A uses such a device to secure a crime scene within a populated neighbourhood. Upon seeing a kid running away with something in her hand, the device misclassifies it as a threat. Hence, it uses lethal force and kills her. It was later determined that she was running scared and carried nothing more than a few toys in her hands.*

[354] As one scholar puts it, "foreseeability will be more relevant with something closer to artificial general intelligence ('AGI'), sometimes called 'strong AI.' However, "AGI is at best many years off and essentially unrelated to existing machine learning technologies." (ibid., 1344.). See also Calo, (2017), 432. He argues that "nothing in the current literature around ML, search, reinforcement learning, or any other aspect of Al points the way toward modelling even the intelligence of a lower mammal in full, let alone human intelligence."

[355] HRW and IHRC, (2015).

[356] United Nations Security Council, (2021).

[357] HRW and IHRC, (2014).

[358] HRW and IHRC, (2015), 25.

If one takes the definition that prevailing positions use, the question is whether the officer was in control of the autonomous device's behaviour. And the answer is no. The robot's autonomy excludes human control (1.a.i,). The device itself fulfils the conditions to say that someone is criminally liable. Hence, one cannot attribute its behaviour to anyone behind. In this vein, HRW would argue that it is unclear "who could be liable when an autonomous machine makes life-and-death determinations without meaningful human intervention."[359]

The model here proposed suggests a different assessment. Take the officer and ask, no whether the device decided to harm, but whether she could have understood, considering her background knowledge and perception of the circumstances, whether the robot would end up killing a person (1.b.ii.).

One could argue, for instance, that she had enough experience to realise that populated areas are prone to erroneous use of force. In this vein, she could use the motivations stemming from applicable ethical guidelines in her reasoning. Consider, for instance, the General Provisions enshrined in the Basic Principles.[360] Under these, law enforcement agencies should restrain the application of means capable of causing death to persons. Upon perceiving that the crime scene was within a populated area, she could thus have been expected to form a belief that the robotic device would end up misclassifying a bystander and engaging it with lethal force.

Again, questions of whether the use of the LAWS reflect a lack of sufficient concern, or whether more care would have put her in a position to avoid harm, present no particular problem here (1.b.ii-iii.). Depending on the circumstances, a court could perfectly maintain that the use of an autonomous device in such circumstances despite the guidelines enshrined in, among others, the Basic Principles, demonstrates a lack of concern for the interest of others. And it seems plain that other, less-autonomous alternatives, would have put her in a position to control the use of lethal force.

Undoubtedly, the question for attributing criminal liability in scenarios of excessive use of force mutates when a robotic is involved. And one could certainly expect practical problems and evidentiary hurdles. However, that does not mean that no question at all is possible, nor that attribution is irremediably frustrated. What it means is that criminal liability must be attributed on different grounds.

[359] HRW and IHRC, (2014).

[360] See, e.g., General Provision 5 (Congress, (1990).

### CAVs

Here the target is a remark included in the first documents an international organization has issued on the problem of criminal liability. Contrary to HRW, the CoE presents the case of CAVs as a paradigm of the problems that criminal attribution might introduce. Such a prospect was introduced in both a thematic session within CAHAI and in a concept paper developed within the Council. The latter piece stated that "if AI takes the place in the driver's seat there could be a responsibility gap."[361] CAHAI is even blunter in indicating that:

> *... a human user cannot be held criminally responsible for offences which arise solely out of the dynamic driving task in principle, e.g. dangerous driving, speeding and manslaughter*.[362]

Read in light of those statements, it seems that CAVs would present a gap insofar as accidents stem from the AI's driving dynamics. Rendering control to an AI-based driver would impede the attribution of criminal responsibility back to the person in the driving seat (1.a.). However, the model here introduced shows that that is not necessarily the case.

The reason is that one no need to ask who was driving to determine who is responsible. CoE's and CAHAI's approach consider it a decisive factor and, in so doing, side-line the relevant grounds to attribute responsibility. The appropriate question is whether, in light of the circumstances and the driver's or developer's background beliefs, she failed to grasp that AI would pose a risk of an accident (1.b.i.).

It is thus important to flesh out the standards of care an agent is obliged to follow. Did she know that giving autonomy to the car could present the risk of accident? What do ethical guidelines say about her obligation to monitor the vehicle? What are the exclusionary reasons flowing from those standards? Those are complex questions, and the answer is largely contextual. However, none of them point to autonomy as a relevant factor. That the car is steered by AI, or not, is not a sufficient reason to maintain that there is an attribution gap. In this vein, one can continue to determine whether such a failure says anything about the person's carelessness and whether showing more concern would have led her to be able to control the vehicle. If the answer is affirmative, liability attaches to the human "behind" —even if she was not driving.

[361] CPDC, (2018), 4.

[362] Gless, (2020).

Blurred attribution?

Does any of the examples above lead to scenarios of unforeseeable prohibitions (2.1.) does the state end up impaired from deeming someone liable for injuries or life-privations (2.3)? Is the mental link between the driver and the offence severed as soon as she renders control to the machine (2.2)?

The answer is no. Saying that someone behind an autonomous device is liable establish a mental link between her behaviour and the wrongdoing, and such an appraisal does not render criminal prohibitions less foreseeable. On the contrary, the act of delegating control provides sufficient grounds to appraise whether she failed to show sufficient concern for others. Furthermore, there is no reason to refrain from using the criminal apparatus in a manner that frustrates duties of redress. Insofar as a domestic system enshrines these requirements, it is also capable of attributing robot's harm back to the human developer or user.

Here again, practical problems might arise in gathering evidence or determining who should bear responsibility. However, those problems offer no reason to consider that robotics and AI would impede attributing their harm back to developers or users.

### *ii. Does autonomy impede criminal attribution?*

Like unforeseeable events, third-party interventions exclude causality. They impede saying that an event E is the upshot of the behaviour of an agent A. Consequently, they might also impede the realisation of the elements in the definition proposed. If the result is the consequence of someone else's behaviour, one can argue that showing more concern for others while forming belief would have made no difference (2.b.iii.) What kind of autonomy blocks attribution? How autonomous should that intervening party be to impede attributing harm to the first agent?

In Common Law systems, the standard is that an agent is not liable for third-party interventions that are of a free, deliberate, and informed nature—be they foreseeable or not. [363] The standard is similar in Continental Law.[364] The distinction is thus between an intervening cause that acts deliberately and with freedom and one that does not.

What it means for a cause to be free, informed, and deliberate remains an undefined matter and differs across jurisdictions. Notwithstanding such debates, the standard is high. Indeed, it is generally understood that acts of children, insane persons, or incapacitated people thanks to

[363] Ashworth, (2009). Ch. 4, 15-19; Hart and Honoré, (1985), 340-47.

[364] See e.g., Roxin, (1994). § 11, 68-71; With references to France, Russia and Germany, see Heller and Dubber, (2011), 209 - 05, 414-55, 88-531, .

drink or coercion do not impede attributing harm. The intervening agent should, at least, be able to know the kind of action it is performing and whether they are prohibited or not.[365]

Do robots exhibit that kind of capabilities? Recall that robots are autonomous as long as they can sense the environment and decide upon the best alternative to achieve a pre-defined goal.[366] Even the most advanced forms of AI are as good as the data they are feed with and the model to process that data.[367] Indeed, robots display "decisional autonomy" because they process inputs and determine outputs that are not entirely dependent on the programming.[368] However, they cannot grasp norms nor understand whether their behaviour is prohibited or not.[369]

Therefore, the device does not exhibit deliberate nor free behaviour. It cannot choose its own goals nor pick means outside the parameters outlined in its design.  Most advanced robots are more akin to the case of a child or an incapacitated person who cannot correctly represent the world and the wrongfulness of her actions. As Thomas Weigend and Sabine Gless put it, robots cannot be conscious of their freedom nor grasp the concepts of having rights and obligations.[370] In this sense, they cannot exhibit the kind of informed behaviour that the law demands to block causality.[371]

### C. Complexity and blurred attribution

It is now convenient to turn to the last bone of contention of attribution gap theories: complexity. Here again an example (Sub-section i.) is followed by a brief account of why the "many hands problem" pose no hurdle for the attribution of robotic wrongdoings.

[365] Hart and Honoré, (1985), 326-29.

[366] Cf. Lagioia and Sartor, (2020), 448-51.

[367] Rahwan et al., (2019).

[368] Fjelland, (2020).

[369] Cf. Lagioia and Sartor, (2020), 448-51.

[370] Gless, Silverman, and Weigend, (2016).

[371] Against, Law Reform Committee, (2021), 2. The authors observe that *with fully-automated, "human-out-of-the-loop" systems (where the RAI system, within set parameters, makes and executes decisions without any human input or interaction), there may be no identifiable human user involved. This raises questions as to who, if anyone, should be held responsible for any harm caused, and on what basis.* However, they fail to clarify *why* such an autonomy would impede attribution.

### i. *Complexity as a case of blurred attribution*

Uber ADS accident

Consider the example of the first criminal procedure involving a robot. When a Volvo steered by an ADS crashed against and killed a pedestrian, the NTSB issued a report that epitomises complexity concerns.[372]

Consider the ADS. That functionality entailed no less than three systems, each with various software and hardware components.[373] The NTSB determined that the company failed to manage the limitations of its driving system, including its inability to "correctly classify and predict the path of the pedestrian crossing the road midblock."[374]

However, there is no reason to stop there. The report further acknowledges that developers within the company decided to deactivate the car's embedded forward collision warning and automatic emergency braking system without replacing its full capabilities.[375] Had these been in place, it was determined, the accident could have been prevented.[376] Moreover, the vehicle's operator's "prolonged visual distraction" led to "her failure to detect the pedestrian in time to avoid the collision."[377]

Attribution gap theories would point to the complexity of the imbroglio as the main hurdle in identifying a fitting candidate. However, the fact that there are many people involved does not create an attribution gap. Even if it is hard to find out who causally contributed to what extent in cases of many hands as this one, and even if each contribution is small, that just means that it is fitting to hold accountable very many individuals. It gives no reason to assume an attribution gap.[378]

In this sense, that there are many hands involved does not impede to determine who, among them, failed to form a belief upon the risks (2.b.i.), nor whether these failures stem from her character (2.b.ii.). Certainly, one could argue that, in long chains, it is hard to ask the nodes to foresee the impacts of their contributions. Yet, the facts of the case provide enough ground to determine that those who eliminated the redundancies, for instance, had reasons to believe that they were creating a risk of harm. Indeed, eliminating those redundancies, or deploying in the street a vehicle with a software that cannot properly classify pedestrians, seems enough to

[372] National Transportation Safety Board, (2020).

[373] Ibid., 8-11.

[374] Ibid., 57.

[375] Ibid.

[376] Ibid., 31.

[377] Ibid.

[378] Köhler, Roughley, and Sauer, (2017), 58-60.

determine which nodes should bear responsibility. Again, the counter-factual judgement pose no problem here. As the NTSB's tests determined, keeping the car's emergency brake, for instance, would have reduced the possibilities of fatally harming the pedestrian.[379]

Blurred attribution?

Seen through the lens of the model proposed, there are no reasons to think that the complexity behind the robot makes it illegitimate to impose criminal liability. It is not only possible to find a fitting candidate. It is also legitimate to impose blame on those who demonstrated insufficient concern by failing to form a belief upon the risks that such a car was posing if left in the streets. Again, there is no reason to think that criminal prohibitions are less foreseeable, nor that the complexity of the device would blur the mental link. Nor are there grounds for a state to refrain from mobilising its criminal apparatus. Human rights standards, as developed within the ECtHR, allow doing so. In this sense, there are no motives to consider that complex devices, as the Uber case epitomises, blurs criminal attribution.

### *ii. Does complexity impede criminal attribution?*

Closely seen, complexity is a shortcut for two issues in attributing robot's harm. The first of them is that such an imbroglio makes it impossible to find a cause that is substantial enough to merit attribution. Whenever harm is the consequence of many insignificant contributions, the network itself and not one or some of the nodes are the cause. Criminal law's personalised blame demands identifying that specific node, but that is impossible due to the "many hands" and "many things" involved. Hence, it is impossible to find a fitting candidate for criminal liability. That is what Ryan Abbot calls the problem of "legal irreducibility,"[380] and Mireille Hildebrandt discusses under the notion of "making a difference." [381] No node makes a substantial contribution —or makes the difference— so no one can be blamed.

The fallacy behind Abbot's and Hildebrandt's account is that, once they acknowledge the complexity of some robotic devices, they conclude that harms are necessarily mere accidents or the responsibility of a network that cannot receive criminal blame. However, difficulties in attributing criminal liability do not imply that such an attribution is blocked. On the contrary, it calls for a discussion of how to unravel that complexity. The model offers a way of unravelling such an imbroglio. Assuming that there is an unjustified risk of harm ─in the case at stake, a risk to life─ one can single out the nodes which, through a failure of grasping those risks, showed insufficient concern for the interests of others. Certainly, the problem of sub-

[379] National Transportation Safety Board, (2020), 41-42.

[380] Abbott, (2020), 114.

[381]Hildebrandt, (2008), 170.

stantiality of the cause kicks back as one of what it means to show sufficient concern. However, that is a matter of how robotic systems are understood in each society and has nothing to do with the fact that there are many hands.

Mireille Hildebrandt introduces the second issue. In her account, not only finding a node that makes a difference becomes intractable. Criminal attribution is also diffused insofar as none of the nodes within the network might *know* the harmful outcome. Whenever many hands bring in small contributions to putting a robot in place —imagine many individuals writing discrete lines of code— it seems that any harmful outcome is outside the range of what contributors are aware of.[382]

The second flaw here is that it ignores that criminal attribution is not necessarily based on actually knowing an outcome. Indeed, the *failure* to become aware of risks also attracts criminal liability. That certainly raises important issues in attributing criminal liability for machine behaviour. Under which circumstances has the node "behind" the robot failed to form an adequate awareness of the risk of harm? What are the criteria for ascribing that lack of knowledge, and how do emerging machines challenge them? These are important questions that, nonetheless, remain obscured by pointing to complexity as an obstacle. Even if they are challenging, nothing in the robot's complexity makes them intractable.

## D. Conclusion

This Chapter saw the model introduced in Chapter 4 in action. It showed that attribution puzzles dissipate as soon as one couches an adequate vision of criminal attribution.  It is correct that the intervention of unforeseeable events or deliberate third parties negate attribution and, thus, impede criminal liability. However, it is wrong to say that *robotic unpredictability* and *robotic autonomy* do so. None of them resembles the kind of situations introducing a failure of accountability. Nor does the fact that robots presuppose a complex network of people and machines.

The Chapter's takeaway is thus the following: the reason why robots block attribution is not to be sought in their complexity, autonomy or unpredictability. Yes, these devices might involve a network of people and might exhibit unforeseeable behaviour. However, insofar as domestic laws enshrine liability for failures to display enough concern, these features would not impede fulfilling human rights-based duties of redress.

[382] Ibid.

# 6 Reasons for concern: when robots blur criminal accountability

The previous Chapter contested the claims that attribution gap theories have been making. If one focuses on the elements of the proposed model, it becomes arguable that neither robotic unpredictability nor autonomy or complexity suffice to block the attribution. Is it thus reasonable to conclude that robots pose no problem for duties of redress? That there is no disruption of the legal framework?[383]

The answer is no. Attribution gap theories miss the target, but that does not mean that there is no target at all. With methodologies closer to that of Ryan Calo[384] or Andreas Matthias,[385] these theories end up focusing on a fixed set of characteristics ─the oft-mentioned trio of complexity, unpredictability and autonomy. The upshot is relegating what is particular of the inadvertent human "behind" the robot. However, as soon as one brings the human to the forefront, the kind of disruption that AI and robotics introduce becomes more perceivable.

If the previous Chapter argued that the predominant puzzle is not as insurmountable as it appears, this Chapter depicts two fundamentals yet ignored problems. Section A introduces the first of them. It argues that, as AI and ML techniques merge with human reasoning, criminal blame becomes unresponsive to user's display of sufficient concern. In turn, Section B speculates on the overstretching of expectations of reasonableness on individuals through the supply chain. If some techniques alienate users from their beliefs, others bring the "developer" forward to act in scenarios where their perception and knowledge is limited. Section C concludes.

## A. First case of blurred attribution: non-intuitive AI and the failure of sufficient concern

This section introduces ML-based decision-making as a case of the second type of blurred attribution: a rupture of the "mental link" that human rights standards demand.[386] The problem has to do not with *the device's* autonomy or unpredictability but with how decision-assistance technologies challenge human rationality. Instead of the metal-cladded robots or the fully autonomous weapon making life-and-dead decisions, these technologies operate by making recommendations to a user. However, their discreteness should not mislead. It is submitted here that they introduce a layer of non-intuitive and inscrutable reasoning between

[383] For a definition of "legal disruption," see Liu et al., (2020), 206-07.

[384] Calo, (2015), 532-48.

[385] Matthias, (2004), 176-81.

[386] For the explanation of the standard, see Chapter 4.

users and the consequences of their actions, such that her behaviour falls short of reflecting the kind of mental link that criminal liability requires.

The argument follows three steps. Sub-section i. introduces ML's non-intuitiveness and inscrutability as an alteration of user's possibilities of action. Sub-section ii. shows how such affordances turn decision-making into an exercise that is unresponsive to the kind of sufficient concern that criminal attribution presupposes. Sub-section iii. argues that such a situation introduces cases of the second type of technologically blurred attribution.

### *i. ML decision-making as a non-intuitive layer*

Starting with Matthias's seminal piece,[387] scholarly and policy debates about criminal attribution often have AI, and particularly ML, as its target.[388] What is so particular of this technique, however, remains clouded behind terms like secrecy[389] or opaqueness.[390] Terminological differences aside, however, what is of interest here is the nonintuitive character of ML's output, often combined with large amounts of data.[391]

What does it mean that ML-techniques are nonintuitive? In a nutshell, it means that machine's reasoning does not square with human reasoning. Take a system recommending an officer whether the situation is one requiring lethal force. How does the human think? She might use her senses to see if the target possesses a weapon. She might look at the target's hands and see whether there is something that matches her understanding of lethal tools. She might also use her experience to see if it is imminent that the target will use the weapon. Now, take the ML decision-assistance system. It might look, not at the weapons or the target's movements, but it might take gait patterns, previous behaviour, and ethnic traits. It might then contrast those discrete data points with terabytes of data and build a model to make a prediction on whether that person is likely to use force. And then, comes the recommendation.

Going back to the notion of *nonintuitiveness,* what it thus mean is that the human user is essentially unable to "weave a sensible story to account for the statistical relationship in the

[387] Matthias, (2004).

[388] For definitions of these terms, see Chapter 1.

[389] See e.g., SIENNA project, (2019); Pasquale, (2011).

[390] See e.g., Burrell, (2016).

[391] Selbst and Barocas, (2018), 1096-98. However, opacity and nonintuitiveness are connected problems. See, e.g., Bygrave who argues that "The problem of opacity reflects not simply shortfalls in humans' computer programming skills *but the fact that the decisional processes involved do not closely emulate the logic of human thought processes*" (Bygrave, (2020), 7.).

model" that ML-techniques build as they interpret training data.[392] The reason is that it defies human sense about the relevance of certain variables.[393] It supposes, in the words of Paul Ohm, that "we are embarking on the age of the impossible-to-understand reason" where decisions are made on the basis of odd correlations.[394]

Such a feature would not be problematic if those odd correlations were always wrong. In facing Lee Sedol, the world's best Go player, AlphaGo's 37th movement was so unusual that seemed like a mistake.[395] However, the move turned the course of the game and AlphaGo went to win.[396] And yet, ML-techniques also fail. They might reduce the chances of errors, but things will continue to go wrong from time to time.[397] How to spot the difference between right decisions and mistakes if none of the solutions square with human reasoning? That is the problem coined as AI's "nonintuitiveness."

Why not making systems more explainable?[398] Why would it not suffice for a developer to follow standards, like the emerging IEEE P7001 on transparency in autonomous systems?[399] That would certainly improve the situation of the officer –who would get an explanation of which elements were relevant (i.e., ethnicity, gait pattern) and what model did the machine use to render a prediction.[400] However, that does not suffice to make the system intuitive. The problem here is not one of making the basis for deciding less obscure, but that it fails to square human understanding of phenomena. More or less explainable, machines and humans reason differently. And there is no principled way to say who is wrong.

Nor is going through the input data a feasible solution. One could imagine an officer who tries to go also through information on ethnicity and gait patterns to see if she can reach the same

[392] Selbst and Barocas, (2018), 1097. See also Bygrave, who departs Ulrich Beck's *Wissenssouveränität* as a framework to pin down how ML-based decision making systems might impair our cognitive abilities (Bygrave, (2020), 9-10.).

[393] Selbst and Barocas, (2018), 1095.

[394] Ohm, (2012), 1309, 18.

[395] See Metz, (2016). Who narrates how a commentator, himself a high-ranked player, thought it was a mistake.

[396] Ibid.

[397] Selbst, (2020), 1321.

[398] The field of explainable AI (XAI) is precisely concerned with the effort to provide explanations about the mechanisms and decisions of AI systems. And one could concede that those efforts might render technologies whose reasoning a user could contest. However, the fact that those efforts exists grounds rather than deflates the argument herein made. Claims that AI must be *made explainable* confirm rather than contest its lack of intuitiveness (See Graaf and Malle, (2017).

[399] Draft Standard for Transparency of Autonomous Systems, IEEE P7001 (IEEE, PAR approval:  2016-12-07).

[400] For instance, under IEEE P7001 a device meeting level 4 of transparency "[…] should be equipped with a "why did you just do that?" function which, when activated, provides the user with an explanation of its previous action, either as displayed or spoken text" (Winfield et al., (2021). However, giving the why does not guarantee that such a why matches human reasoning, nor that the user is in a position to contest it.

prediction that the machine reached. However, that falls short of a feasible solution. The point of ML is precisely its application to those problems where encoding and explicit logic functions very poorly.[401] Those scenarios often involve situations with many terabytes of data, where the number of possible features rapidly grows beyond what human reasoning could ever grasp.[402] Humans fail exactly where ML thrills: in processing trillions of data examples and thousands of properties.

The claim for intuitive relationships is not a demand for disclosure or accessible explanations. It means one that the machine should rely on reasoning that comports with intuitive human understanding. The notion of "face validity" explains this feature. Used in psychology to assess psychological tests, face validity is "the appropriateness, sensibility, or relevance of the test and its items as they appear to the persons answering the test."[403] What it thus captures is that some measure becomes credible insofar as it squares with the agent's understanding of a phenomenon. Making a device more or less explainable does not necessarily mean that the reasons match human understanding or will be perceived as relevant. Yet, how to know whether those odd reasons are valid escapes human capabilities. There is no principled way of setting differences of criterion between an individual and ML's applications. That is the point of nonintuitiveness.[404]

Robots might not poison the household —at least not with their developers being shielded from liability— but one should not ignore how ML-based decision-making techniques add a layer of unintuitive reasoning between a person and the consequences of her actions; such a layer being based on information that the human user cannot fathom.

### *ii. The downfall of failures to belief and sufficient concern*

Why are decision-making tools a reason to worry? How would they affect the attribution of criminal liability? Take the case of medicine, where AI is increasingly pitched as capable of finding correlations and predicting things that even well-trained humans cannot.[405]

Imagine an algorithm fed with data from insurance claims, sensors and electronic medical records to provide personalised treatment recommendations.[406] What if the technique recommends a treatment that, once followed by a well-trained doctor, ends up killing the patient?

[401] Burrell, (2016), 6.

[402] Ibid., 9. As Domingos rightly notes, intuition fails at high dimensions (Domingos, (2012), 82-83.).

[403] Holden, (2010). Also using the concept, Selbst and Barocas, (2018), 1097.

[404] Some authors, however, might understand that feature as part of the problem of explainability. See Bygrave, (2020), 7-8.

[405] For reviews of the state of art, see Price, (2018); Selbst, (2020), 1335-36.

[406] See, e.g., Elias et al., (2015).

Does this example sound as utterly fictional? The reality shows a different picture. Decision-making applications are already under current development and introduction. [407] And yes, they make mistakes. According to some documents, IBM's Watson for cancer diagnosis has already come out with odd suggestions, like recommending for a cancer patient with severe bleeding to be given a drug that could cause it to worsen.[408]

As things go wrong, could one attribute an offence to the expert who followed ML's odd advice? Take the first element of the definition: is it possible to argue that, in light of her background knowledge and perception of the circumstances, she failed to form a belief that such an odd treatment would end up killing the patient (1.b.i.)?  A court will be right to answer in the negative.

Recall that the definitions seeks to determine whether a person, in view of her background training, displayed diligence in the process of belief formation.[409] Now, it is true that the agent in the example above decided upon a nonintuitive machine assessment. Yet, as AI-techniques could identify patterns her reasoning would fall short of spotting, she had all the reasons to doubt her own assessment. All she had was an unintuitive output. Whether it was right or wrong is unresponsive to her degree of diligence.[410]

The corollary is that it is impossible to know if she would have avoided the result by showing more diligence, as the third ingredient of the definition recommends (1.3.iii). ML decision-making severs the connection between beliefs and the process of their formation, such that a court cannot easily establish whether more or less diligence would have put her in a position to avoid harm. Imagine she disregarded the machine assessment and followed her own. Is her display of care relevant? What if things would have gone wrong as a consequence of ignoring the machine's advice? One cannot but argue that sge might have also ended up falling short of a forming an adequate belief.

Again, the problem is not the lack of standards of care, as some authors seem to point out.[411] There are already duties for medicine, surgery, driving or data security. What changes is how people make decisions once AI enters the scene and inform those decision-making processes. It in clouding those decisions that AI complicates the assessment that an agent failed to form a belief upon the risks that she was creating.

[407] See Price, (2018).

[408] See Chen, (2018).

[409] See Chapter 5.

[410] Selbst, (2020), 1331.

[411] See Gless, (2016); Gless, Silverman, and Weigend, (2016). In the general literature on liability, see Rachum-Twaig, (2020).

Asking whether such a failure is reflective of her character ─as 1.b.ii. would recommend─ is no less troublesome. The failure to form a belief says nothing about her lack of concern. The mistake instead is the upshot of an erroneous prediction that one must include in the calculation as AI techniques join human decision-making.

One could certainly argue that sufficient care would have demanded to confirm the prediction with a different device. However, such an assessment displaces rather than solves the problem. Be it one, or two devices, all she could do is to trust one prediction over the other. Still, however, the harm cannot be traced back to her failure to care for other's interests. When ML-based techniques make a recommendation, the human would often fall short of knowing on which side she stands in following or disregarding its advice.[412] She might be able to understand the output, and even map it to certain intuitions. Still, there is no principled way for settling the difference between her assessment and that of the machine.

Certainly, most of the affordances that these devices introduce are yet to be fully recognised and acted upon.[413] Still, one can build reasonable conjectures considering both possible and actual applications. LAWS that, even if not engaging a target, recommend to a human officer whether to use force or not might be a case in point. Similar considerations could also apply to an AI-based interrogator, which, upon mining data in thousands of questionings and testing different strategies, might make recommendations ─like using specific patterns of voice─ that end up causing great suffering equivalent to ill-treatment.[414] What is particularly dire in these cases is that AI builds its own categories unintuitively ─ "threat requiring lethal force"─[415] or acts upon the target it aims at predicting. In this vein, one cannot even establish whether the agent was right in following the device's advice in engaging it. Nor it is possible to assess whether the interrogatee would have been better without AI's intervention.

Notice the difference with cases where the human deploys an autonomous or unpredictable robot. In those cases, it is possible to trace the *decision* to deploy or develop back to a human agent. One could peep into her background beliefs and professional trainings and ask whether she *failed to perceive* that the circumstances where unfit for using a device. The court would find the same ingredients that it would find in any criminal assessment. The challenge that AI as a decision-assistance tool introduces is of a different nature. It intervenes with the *process of making decisions* itself and challenges attribution a fundamental level. Returning to Paul

[412] Selbst, (2020), 1332.

[413] 'Affordances recognised and acted upon' refers to possibilities of action that, once introduced, are also identified and put to use. See Liu et al., (2020), 224.

[414] See Thomasen, (2016).

[415] Cf. the example of data security in Selbst, (2020), 1337.

Ohm's phrase, replacing human reasoning for an "impossible-to-understand reason"[416] allows no judgement on a person's display of concern.[417]

### *iii. A case of technologically blurred attribution*

Such an impact on human decision-making introduces a predicament difficult to overcome without frustrating duties of redress or unjustifiably interfering on defendants' rights.

Imagine a court deems a person liable for failing to grasp the risk that ML was making a mistake. Would that say something about her lack of concern for others' interest? That it is not possible to tip the balance in favour of one option makes clear that, whatever happened, does not speak of her level of diligence. Nor would it serve to say that she had a background belief that AI *sometimes* errs; for the problem is that its nonintuitiviness blurs determining *when* the outcome is erroneous and when it is not.

Deeming her criminally liable despite these concerns makes it a case of attributing harm without any mental link between the offender and the wrongdoing. The latter is the product of a diligent yet technologically obscured belief formation. Hence, putting her under the aegis of criminal law says nothing to her, or to others, about the blameworthiness of her behaviour. She ends up in a position where she is blamed not for her choices or dispositions but for following a generally more accurate yet unintuitive tool. Thus, her human right not to be deemed liable without the establishment of a mental link ends up unavoidably restricted.

Is such a restriction nonetheless legitimate? Is it possible to argue that it is a necessary means to securing criminal redress in the face of emerging technologies? If so, is the defendant left with some alternatives to exonerate herself? Both questions should receive a negative answer.

Duties of redress end up denaturalised if punishment is imposed at all costs. The later are part of human rights obligations to protect, not to coerce.[418] And, as Chapter 2 showed, a framework that meets the right's threshold should not only be effective but also dissuasive and proportionate. These qualifications introduce a demand to impose punishment only insofar as the guilty target can be identified on fair grounds. A system that imposes liability despite the person showing sufficient diligence would betray those requirements. Thus, it would fall short of a necessary mean to secure criminal redress in the face of emerging technologies.

[416] Ohm, (2012), 1318.

[417] Defending a different point, Law Reform Committee, (2021). It argues that *any criminal liability for a harmful act involving an RAI system should be imposed is likely to be a function of: (a) the severity and risk of actual or potential harm inherent in the use of the system in the relevant context; (b) the level of automation of that system; and (c) the degree of human oversight over, and involvement in, the system's decision-making (if any).* It is precisely that possibility of 'oversight' what AI in decision-making blurs.

[418] Lazarus, (2012).

Even if dissuasive, presuming that the defendant failed to show sufficient concern would leave her with little to no chances of exonerating herself. For what could she argue if making further investigations or relying on her own assessment would have not ended up as a display of sufficient concern? If the officer disregards a LAWS advice and ends up omitting her obligation to stop a threat to others' life, she would be liable for failing to form a belief despite what a more accurate tool advised. If, on the contrary, she follows the device's erroneous advice, she would be equally liable for failing to perceive the circumstances in light of her training and experience. And further investigation, if possible, would have not take her farther from imposing an interpretative reasoning on a mathematical process of statistical optimisation.[419] A convincing defence is, at least in theory, not viable.

Holding punishment and deeming the agent as innocent thus seems like the best outcome. However, it is also one that falls short of human rights standards. Here, the problem is that duties of redress demand liability and punishment where behaviours are generally criminalised or entail attacking right-protected statuses. One can also think of introducing less severe offences that would somewhat bear some proportionality to the lack of a mental link.[420] However, if the robot's wrongdoing entails live deprivation, or ill-treatments, the state will be also falling short of its obligations to redress victims.[421]

The result is thus a predicament where solutions will either entail a combination type two ─liability attributed despite no mental link being established─ and type three ─frustration of duties of redress─ blurred attribution, or an autonomous case of the latter kind. Either alternative is unsatisfactory under human rights standards and, thus, can be rightly classified as instances of technologically blurred attribution.[422]

### B. Overstretching the reasonable person

Some voices, like that of Thomas Weigend and Sabine Gless, have rightly point out that the problem of criminal attribution is one of targeting the inadvertent actor "behind" the machine.

[419] Burrell, (2016), 9.

[420] See Law Reform Committee, (2021), 38-42.

[421] Recall that in *Volodina* and *Öneryildiz*, the ECtHR was not satisfied with targeting the offender with a less-severe crime. See Chapter 2.B

[422] Against Dsouza, (2020), 256-60. He argues that *the addition of an autonomous AIT makes little difference to the analysis of the human defendant's knowledge or belief. We simply need to ask whether D performed her conduct (be it programming, using, hacking into or – on a stretched interpretation of the word "conduct" – knowingly owning the AIT) with the knowledge or belief required for the offence. The knowledge or beliefs of the AIT (assuming that AIT actually forms beliefs in the same sense as humans form beliefs) are irrelevant, since it is not the defendant* (256-57). However, the problem is not whether the beliefs are the agent's or the artefact, but to what extent using the later blurs the former's belief formation. The author ignores that point.

For them, however, the question is one of defining standards of reasonableness developers and users can follow.[423] They would point to early examples already featuring in robotic products and AI embedded in different systems.[424] If their explanation follows, a robot is not so different from a bridge or any other product.[425] Attribution failures are nothing more than a upshot of insufficiently recognising the emergence of these standards or, at their best, an effect of their ─transitory─ lack of maturity.[426]

Such an approach certainly indicates an adequate starting point. However, pitching it as the solution is close to question-begging.[427] The problem is rather *to what extent* can standards of reasonableness regulate robots in the same way they regulate other artefacts. The previous section showed that, with the advent of decision-making technologies, those yardsticks fail to reflect a user's sufficient concern. This Section introduces a further problem, this time relevant for developers ─i.e., integrators, designers, coders and cloud service providers─ called to take part in the supply chain "behind" autonomous CPSs. Indistinctly coined as developers to facilitate the exposition, it argues that robots, and particularly CPSs embedding some form of autonomy, overstretch the reasonableness that is expected of them.

Section i. introduces two problems underpinning the case: bringing developers under standards of care that they are not equipped to follow and difficulties to validate autonomous robotics. The reason why this challenges attribution is in turn the focus of Section ii. Section iii. presents a case of the first and third type one blurred attribution: lack of foreseeable prohibitions and frustration of duties of redress.

### *i. Difficulties in validating robotic behaviour and the expansion of reasonableness*

Powerful autonomous systems now share in the physical and e-space that used to be the exclusive realm of humans. The difference between a caged industrial robot, or autonomous subways that have been in operation for more than forty years, on the one hand, and a LAWS or a robotic interrogator, on the other, is that the latter do not operate in segregated environments.[428]

[423] See Gless, Silverman, and Weigend, (2016). Gless, (2016), 5. (who argues that the "control dilemma" of the car driver is basically a "risk allocation dilemma" of the legislator); Gurney, (2015).

[424] See, for instance, ISO 13482:2014.

[425] Gless, Silverman, and Weigend, (2016), 427.

[426] See Leroux and Labruto, (2012), 51.

[427] For a similar argument, see Santoni de Sio and Mecacci, (2021).

[428] Fisher et al., (2021), 8.

In this scenario, it is generally considered that, as CPSs take over tasks, responsibilities end up transferred the machine.[429] Even if accurate, however, such an affirmation conceal that a great deal of these choices are rather transferred to developers. In setting the boundaries of robots and validating their decisions, the developer is leaving their ambits to become a sort of caregiver, interrogators and law enforcement officer. This is a new and nor fully acted upon affordance. Yet, it has remained surprisingly unrecognised.

Consider the following example:

> *[…] an assistive robot [is] deployed to aid a human in the inspection of nuclear facility. This robot may be responsible for warning the human of any potential radiation exposure based on their proximity to, and strength of, the radiation. However, to perform this task the robot has to constantly follow the human around, which may eventually annoy or irritate them. If the human asks the robot to stop following it, should the robot comply with the request? In making such contentious and critical decision, how should the robot explain itself if requested? And, in all these cases, what verification will be required?*[430]

As autonomous CPS enter different ambits of life, decisions like these are increasingly less in the hands of a rescuer and more in those of developers. The latter are who define whether the robot should comply with human request. However, these decisions are *mediated*. Developers are not in the heat of the moment when deciding whether the assistive device should stop following the humans it is supposed to help. They cannot assess the level of annoyance, or whether it is reasonable *while* performing the activity. On the contrary, their decisional power goes through a machine who is supposed to decide without human intervention. Hence, they must make critical choices while being far retired from the context where their creations operate.[431]

Such a qualification raises a number of questions. Once some autonomy is delegated to the machine, how can one be sure that correct decisions are always made given the information available?[432] What guarantees can be given and how could it be verified? These poisonous questions inevitably lead to the problem of validating and verifying autonomous decisions.

Indeed, as the level of autonomy in the system increases, there is a growing gap between what can be determined in a test and how the device behaves in real-life scenarios. Indeed, different

[429] See Beck, (2016), 140. ("By developing autonomous robots with the ability to learn, we are building machines overtaking responsibilities even on the stage of decision making"`).

[430] Fisher et al., (2021), 21.

[431] See ibid., 19-22.

[432] Ibid.

techniques are irremediably based on imperfect models of the rather rich environment where these devices are supposed to operate. Mathematical models but also digital twins fail to capture all those circumstances. Physical testing, in turn, cannot be replicated. Once tested, either the robot itself (think of learning, for instance) or its environment (temperature or obstacles) will change in manner that outdates the previous assessment. At best, developers could end up with an approximation based on the combination of different methods.[433]

The upshot is a new affordance: their creations take developers to contexts they are unfamiliar with, whereas they lack the tools to verify that their machines will make the right decisions. What is the impact of these new affordances on criminal attribution? How does it blur accountability?

### *ii. Overstretched beliefs*

Consider the following example:

> *A robotic surgeon is programmed to autonomously perform certain cuts and manipulate tissue, particularly in circumstances where its remote human operator loses signal or cannot steer the device. Even if futuristic, one could imagine these CPSs operating in circumstances ─political turmoil─ where a human surgeon needs to act remotely and under the risk of losing connection.[434] Now imagine that, while operating autonomously, the device faces an unexpected body reaction. The machine does not react appropriately and ends up damaging the tissue and causing the patient's death.*

One would be tempted to attribute the damage back to the developers. They were the ones who programmed the device to make decisions in such an environment and, thus, are the ones better suited to answer in case those decisions go wrong. Any attribution, however, will firstly encounter the problem of background beliefs. Was developers' latent knowledge about surgeries, unexpected hazards and tissue manipulation such that one can find the necessary background training? Would their training and expertise allow them to identify the hazard that ended up killing the patient? The likelihood of a negative answer is what makes the case of attribution particularly complicated in these circumstances.

Even assuming that they had the background beliefs necessary to understand the risks, could one say that they failed to perceive them considering that such a perception is rather mediated by validation and verification processes that are limited in capturing all the circumstances? Was it possible for them to "perceive" the hazard at the moment where the device was pro-

[433] Ibid., 12-19.

[434] Cf. O'Sullivan et al., (2019), 2.

grammed? Here again, a positive answer is complicated. While alienating those ─like the surgeon─ who might have had the necessary expertise and perception to attribute her a failure to react appropriately to the risk, CPSs put the focus on a supply chain of whose nodes a court cannot equally say that they failed to display such features. The upshot is a difficulty in demonstrating a failure of belief.

One could contest that developers are liable for assuming an activity despite lacking the necessary expertise.[435] One could also point to the emerging regulations and standards. However, neither of these responses is satisfactory. First, the assessment presupposes that it is legitimate to delegate the activity to a robot. If that were not the case, no problem of attribution would have arisen. Hence, the question is that, considering some activities are legitimately delegated, who should be accountable when things go wrong?

The second objection fails as one zooms into those regulations and standards. Looking into the Medical Devices Regulation,[436] or the proposed AI Act,[437] says little of how to deal with these situations. Apart from few specific requirements,[438] these norms, as well as technical standards,[439] rely on developers' own assessment of risks. [440] They might be even obliged to ponder foreseeable uses along with misuses. The problem, however, is not one of a failure *to assess* risks but of a general limitation in determining what counts as a potential hazard. There is an inherent gap, such that even a compliant developer might fail to form a perception upon all the decisions they are called to take *through* their machine agents. That is precisely were attribution might fail.[441]

[435] Lagioia and Sartor discuss imposing criminal Liability for creating and using some AI systems (Lagioia and Sartor, (2020).

[436] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [2017] OJ L 117, 5.5.2017, p. 1–175.

[437] 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 [2021] COM/2021/206 final) [Hereinafter "AI Act"].

[438] The AI Act is an example in point. Consider the use of data to train AI systems. The Act requires that such a training data shall be "relevant, representative, free of errors and complete." (Article 10 (3) However, the nitty-gritty of what counts as relevant and free of errors data is left undetermined and will largely depend on the governance and management practices chosen (see Article 10 (2).

[439] See, e.g., ISO 13482:2014, 4.

[440] For instance, EU's proposed AI Act often requires setting forth a risk management system for targeted applications (See, e.g., AI Act, Articles 9, 67 (1) or 69).

[441] Cf. Liu and Zawieska, (2017), 324-25.

### *iii. A case of technologically blurred attribution*

An overstretched developer is one who will not grasp the risks of the new activities that, thanks to her creations, she is assuming. Is it legitimate under human right standards to hold her liable for these failures? The answer is arguably no. Why is that the case invites to recall that one is liable for failing to form a belief upon two factors: background beliefs and perception. Insofar as a person lacks the necessary background or cannot perceive the environment, deeming her liable implies making criminal prohibitions unforeseeable.

Recall that the requirement of foreseeability means that a person can know what to do to avoid criminal condemnation; such knowledge being assessed from the point of view of the person at stake. An inadvertent developer, hence, has a right to know what kind of diligence is expected of her to avoid criminal liability. The crux of the argument is that such a right ends up undermined as the developer is deemed liable despite being under circumstances where she has to make decisions lacking expertise and perception of the circumstances. She will be in a position where, no matter how diligent she is, she cannot avoid being blamed for circumstances falling outside her range of expertise and perception.

Is the restriction on her right to a foreseeable prohibition nonetheless appropriate? That is hardly the case. As argued in Section A, duties of redress do not demand punishment at all cost. Insofar as the latter loses its rationality, the former becomes pointless. And that would be the case is someone is deemed liable despite not being able to forecast the prohibition; for criminal law would not be dissuading any behaviour if it cannot establish from the outset what is forbidden. It is closer to developers being surprised with criminal liability than to prohibitions they can follow in shaping their creations.

Holding punishment, even if it is the best alternative to secure developers' rights, is also inadequate. The reason, here also, is a frustration of duties of redress. Victims that, if operated by a surgeon or interrogated by a human, would have found redress, are now left outside the aegis of criminal law just because it was a machine who caused harm. The situation irremediably ends up in a frustration of human rights obligations.

Seen under the proposed definition, such a frustration introduces a case of the third type of technologically blurred attribution. In turn, deeming developers as criminally liable for unforeseeable standards entails a case of the first type and, consequently, also obstructs obligations of redress. Therefore, be it that the developer is punished or spared, accountability will end up blurred.

## C. Conclusion

This Chapter aimed at using the proposed model to unpack some cases of technologically blurred attribution. The Chapter's takeaway is that a focus on robot's unpredictability, complexity and autonomy has led to ignoring two cases of blurred attribution.

The first of them, introduced in Section A, focuses on the users and on how ML-based decision-assistance technologies difficult tracing failures back to displays of insufficient concern. The upshot are instances of the second and third type of technologically blurred attribution. Either blaming or sparing the person "behind" the robot led to unsatisfactory results where both the rights of defendants and duties to redress end up frustrated.

Section B presented a different yet also complex panorama: that of developers. As users are alienated from the reasons behind their decisions, developers are brought to the forefront. However, their lack of expertise, combined with the impossibility to "perceive" future challenges, blurs the attribution of criminal liability. Here again, blaming or sparing the inadvertent defendant ends up in instances of technologically blurred attribution. What these cases highlight is that broadening expectations of care turns criminal prohibitions into unforeseeable commands. No matter what a developer does, she cannot avoid criminal attribution. Such a situation, as well as sparing her, would lead to unacceptable restrictions upon human rights.

Undoubtedly, the list might keep growing and the Chapter's purpose is not to exhaustively present all the cases. Its goal is rather to show how the lens of an improved definition point to cases where calls for fair criminal accountability might end up frustrated. In so doing, it sets the stage for further discussion and, particularly, for devising appropriate regulatory responses. Briefly discussing those responses is the concluding Chapter's goal.

# 7 Concluding thoughts: from upgraded liability to meaningful human control

One scholar has rightly argued that, among the human rights issues AI raises, the "more pernicious" are those that have yet to be identified or articulated; for these are the ones arising from "new affordances rather than directly to AI modelled as a technology." The problem of criminal liability, insofar as it reflects how technology challenges human action, can be rightly put within that cluster of unarticulated human rights issues. This piece was an attempt to start articulating the problem.[442]

The first takeaway is that the attribution of criminal liability is a matter of human rights to the extent that blurred accountability might frustrate the obligation to set forth a legal framework that is dissuasive, proportional and effective.

The second takeaway is that —contrary to most scholars and organisations' view— the engagement of robots in a wrongdoing does not blur accountability because robots are unpredictable, complex or autonomous. These features are certainly problematic. However, they do not impede saying that a user or a developer showed insufficient concern in grasping the risks she was creating by putting such a device in motion. Reaching that conclusion required refining the idea of criminal liability underpinning "attribution gap theories" and replacing it for the idea of "technologically blurred attribution."

The third takeaway —regarding the third question— is that there are indeed persisting reasons of concern. Robots will blur the assessment that a person is criminally liable. That will happen insofar as they interfere with human decision-making. Decision-assistance tools that have the potential to alienate users from their reasons, and autonomous CPSs taking developers closer to unfamiliar ambits are two cases in point.

Now, recall CoE's suggestion of an international framework.[443] What kind of framework would work in light of those persisting issues? How to ensure a fair allocation of criminal liability? One can draw some brief observations from the piece's analysis.[444]

It is worth starting by stressing that targeting the robot with punishment is not a fitting solution.[445] Certainly, it might be effective, particularly when there is no obvious human third-

[442] For an overview of the arguments, see Chapter 1, Section C.

[443] Gless, (2020), 9-11.

[444] For an overview of the thesis's argumentation, see Chapter 1.

party to hold accountable. Yet, it falls short of a dissuasive and proportional framework. It makes criminal liability unresponsive to displays of concern. Would it matter for an officer to be careful if all in all who responds is a LAWS? The answer is no, and the reason is that a framework cannot deter rights-threatening behaviour if the allocation of liability does not depend on showing concern for the interests of others. Deeming the machine liable tells potential offenders that no matter what they do, the robot pays. Hence, it frustrates duties of redress as much as unfairly distributing punishment would.[446]

For similar reasons, vicarious liability would not suffice either.[447] Simply calling a person to account *for* a robot supposed to act on her behalf says nothing about her character or whether she displayed sufficient care.[448] Again, it might simplify the allocation of liability. However, it does so at the cost of severing the mental link between the offender and her acts. Such a conclusion also shows why narrowing down standards of care would not suffice.[449] It might certainly counter the expansion of developers' care. Nonetheless, whether the yardstick for sufficient concern is narrow or broad leaves untouched the problem of how to get personal when AI introduces a layer of non-intuitive reasoning. AI blurs liability insofar as it targets decision-making and simply redefining the benchmark falls short of a solution.

Punishing the robot, vicarious liability and redefining duties of care fall short of a solution. What could work then? One can point general principles. First, domestic laws should revisit those instances where inadvertent agents are deemed liable. It should cover not only instances of injury or death, but situations like ill-treatment and rape. The piece showed that it is possible to establish a mental link in those scenarios. The ball is on states' court, who might have to consider expanding non-intentional offences to ambits where they seem atypical.[450]

However, the key challenge is to ensure that criminal liability "gets personal." Doing so demands shaping liability to reflect how AI modify human capabilities and act accordingly. Autonomous cars, for instance, kick control back to the safety driver when the computer runs into trouble. However, it turns out that humans are bad at continually monitoring a situation without being engaged and then taking over when needed.[451] How could criminal law allocate

[445] See  e.g., Beck, (2016), 141-42. See, e.g., Hallevy, (2013), 178. ("Either we impose criminal liability on AI entities, or we must change the basic definition of criminal liability as it developed over thousands of years, and abandon the traditional understandings of criminal liability").

[446] For a general critique, see Gless, Silverman, and Weigend, (2016), 415-17.

[447] See Diamantis, (2021). Against, Osmani, (2020), 62-63.

[448] Even though securing a fair outcome is among the conditions he poses for vicarious liability, Diamantis fails to determine how could that outcome be fair for criminal accountability (cf. Diamantis, (2021), 8.

[449] For proposals of that kind, see Gless, Silverman, and Weigend, (2016), 430-34.

[450] As an example, see Law Reform Committee, (2021), 30-35.

[451] See Davies, (2017).

responsibility if it ignores limitations of that kind? Hence, the need to reflect not so much on machines, but on how they challenge human capabilities.

How good are people in keeping up while collaborating with an autonomous machine? How to ensure that developers and users can have the capabilities to grasp the risks of developing and using AI and robots? Notions like meaningful human control point in the right direction.[452] Going back to the example of autonomous cars, the newly enacted French framework is also as a good example. It makes a driver liable only if a take-over request is issued and after a specified time for regaining control of the vehicle.[453] Not punishing the robot or making an individual automatically liable for whatever the machine does but making law responsive to her altered capabilities.

The details of a treaty go well beyond this thesis scope. Its main virtue was to articulate one of the human rights problems that robots and AI might introduce. In so doing, it opened up the way for reassessing what is needed of regulators and lawmakers to keep criminal accountability as machines go into untrodden areas. In highways and hospitals, the problem is less of unpredictable and complex entities and more of how humans make decisions while cooperating with machines. Zooming into those capabilities and shaping laws accordingly is the path forward for keeping duties to redress while avoiding a dismantling defendant’s rights.

[452] Santoni de Sio and Mecacci, (2021), 4.1.

[453] Law no. 2019-486 of 22 May 2019 on the Growth and Transformation of Companies, Art. 125 modifying ordinance No. 2016-1057 on the Testing of Vehicles with Delegation of Driving in Public Roads. See also Gless, (2020), 7.

# Table of references

## A. Bibliography

(2011), Panel of Experts established pursuant to resolution 1973. "Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)." United Nations Security Council, 2021.

(CAHAI), Ad Hoc Committee on Artificial Intelligence. "Feasibility Study." Strasbourg, 2020.

(Rapporteur), Mady Delvaux. "Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(Inl)." European Parliament Committee on Legal Affairs, 2017.

Abbott, Ryan. *The Reasonable Robot : Artificial Intelligence and the Law*. Cambridge: Cambridge University Press, 2020.

———. "The Reasonable Robot: Artificial Intelligence and the Law." *European intellectual property review* 43, no. 4 (2020): 276-77.

Alexander, Larry, Kimberly Kessler Ferzan, and Stephen J. Morse. *Crime and Culpability : A Theory of Criminal Law*. Cambridge, U.K. ;,New York: Cambridge University Press, 2009.

Andenæs, Johs. *Punishment and Deterrence.* Ann Arbor: University of Michigan Press, 1974.

Ashworth, Andrew. *Principles of Criminal Law.* 6th ed. ed. Oxford: Oxford University Press, 2009.

Bandes, Susan. "Is It Immoral to Punish the Heedless and Clueless? A Comment on Alexander, Ferzan and Morse: Crime and Culpability." *Law and philosophy* 29, no. 4 (2010): 433-53.

Basch, Fernando Felipe. "The Doctrine of the Inter-American Court of Human Rights Regarding States' Duty to Punish Human Rights Violations and Its Dangers." *American University international law review* 23, no. 1 (2008): 195.

Beck, Susanne. "Intelligent Agents and Criminal Law—Negligence, Diffusion of Liability and Electronic Personhood." *Robotics and Autonomous Systems* 86 (2016): 138-43.

Bernhard, Julian, Patrick Hart, Amit Sahu, Christoph Schöller, and Michell Guzman Cancimance. "Risk-Based Safety Envelopes for Autonomous Vehicles under Perception Uncertainty." (2021).

Bertolini, Andrea. "Artificial Intelligence and Civil Liability - Study Requested by the Juri Committee." edited by European Parliament Committee on Legal Affairs. Brussels, 2020.

———. "Robotic Prostheses as Products Enhancing the Rights of People with Disabilities. Reconsidering the Structure of Liability Rules." *International review of law, computers & technology* 29, no. 2-3 (2015): 116-36.

Bhat, P. Ishwara. "Doctrinal Legal Research as a Means of Synthesizing Facts, Thoughts, and Legal Principles." Delhi: Delhi: Oxford University Press, 2020.

Bien, Zeungnam, Dae-Jin Kim, Myung-Jin Chung, Dong-Soo Kwon, and Pyung-Hun Chang. "Development of a Wheelchair-Based Rehabilitation Robotic System (Kares Ii) with Various Human-Robot Interaction Interfaces for the Disabled." 902-07 vol.2: IEEE, 2003.

Board, National Transportation Safety. "Collision between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018." Washington, DC: National Transportation Safety Board, 2020.

Borrelli, Pablo, John Ly, Reza Kaboteh, Johannes Ulén, Olof Enqvist, Elin Trägårdh, and Lars Edenbrandt. "Ai-Based Detection of Lung Lesions in [18f]Fdg Pet-Ct from Lung Cancer Patients." *EJNMMI Phys* 8, no. 1 (2021): 32-32.

Bratman, Michael E. *Intention, Plans, and Practical Reason.* Cambridge, Mass: Harvard University Press, 1987.

Burrell, Jenna. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big data & society* 3, no. 1 (2016): 205395171562251.

Bygrave, Lee A. "Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions." In *Cambridge Handbook of Information Technology, Life Sciences and Human Rights*, edited by Marcello Ienca, Roberto Andorno, E. Stefanini, L. Liguori and O. Pollicino. Cambridge, 2020.

Calo, Ryan. "Artificial Intelligence Policy: A Primer and Roadmap." *U.C. Davis law review* 51, no. 2 (2017): 399.

———. "Is the Law Ready for Driverless Cars?". *Communications of the ACM* 61, no. 5 (2018): 34-36.

———. "Robotics and the Lessons of Cyberlaw." *California law review* 103, no. 3 (2015): 513-63.

Chalfin, Aaron, and Justin McCrary. "Criminal Deterrence: A Review of the Literature." *Journal of economic literature* 55, no. 1 (2017): 5-48.

Chen, Angela. "Ibm's Watson Gave Unsafe Recommendations for Treating Cancer." *The Verge* (2018). doi:https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science.

Chengeta, Thompson. "Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law." *Denver journal of international law and policy* 45, no. 1 (2020): 1.

"Collision between Vehicle Controlled by Developmental Automated Driving System and Pedestrian Tempe, Arizona March 18, 2018." Washington, D.C., United States of America: National Transportation Safety Board, 2019.

Committee, Singapore Academy of Law Law Reform. "Report on Criminal Liability, Robotics and Ai Systems." In *Impact of Robotics and Artificial Intelligence on the Law*, edited by Simon Constantine. Singapore: Law Reform Committee, 2021.

Compton, Kristin. "Da Vinci Robotic Surgery Lawsuits." *Drugwatch* (2021). doi:https://www.drugwatch.com/davinci-surgery/lawsuits/.

Congress, Eighth United Nations. "Basic Principles on the Use of Force and Firearms by Law Enforcement Officials." edited by Eighth United Nations Congress on the Prevention of Crime and the Treatment of Offenders. Havana, 1990.

Connor, Martin J., Prokar Dasgupta, Hashim U. Ahmed, and Asif Raza. "Autonomous Surgery in the Era of Robotic Urology: Friend or Foe of the Future Surgeon?". *Nature Reviews Urology* 17, no. 11 (2020/11/01 2020): 643-49.

CPDC. "Concept Paper: Artificial Intelligence and Criminal Law Responsibility in Council of Europe Member States - the Case of Automated Vehicles." Strasbourg, 2018.

Crisman, Jill D., Michael E. Cleary, and Juan Carlos Rojas. "The Deictically Controlled Wheelchair." *Image and vision computing* 16, no. 4 (1998): 235-49.

Crofts, Penny. *Wickedness and Crime: Laws of Homicide and Malice.* London: Routledge, 2013. doi:https://doi.org/10.4324/9780203409787.

Danaher, John. "Robots, Law and the Retribution Gap." *Ethics and information technology* 18, no. 4 (2016): 299-309.

Davies, Alex. "The Very Human Problem Blocking the Path to Self-Driving Cars." (2017). doi:https://www.wired.com/2017/01/human-problem-blocking-path-self-driving-cars/.

de Casadevante Romani, Carlos Fernández. "The Existence of Common Elements in the Different Definitions of Victim." 89-96. Berlin, Heidelberg: Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

Diamantis, Mihailis. "Vicarious Liability for Ai." In *Cambridge Handbook of Ai and Law*, edited by Kristin Johnson and Carla Reyes. Forthcoming: Cambridge University Press, 2021.

Domingos, Pedro. "A Few Useful Things to Know About Machine Learning." *Communications of the ACM* 55, no. 10 (2012): 78-87.

Donahoe, Eileen. "So Software Has Eaten the World: What Does It Mean for Human Rights, Security & Governance?" *Just Security* (2016). doi:https://www.justsecurity.org/30046/software-eaten-world-human-rights-security-governance/.

Dsouza, Mark. "Don't Panic. Artificial Intelligence and Criminal Law." Chap. 11 In *Artificial Intelligence and the Law: Cybercrime and Criminal Liability*, edited by Dennis J. Baker and Paul H. Robinson, 247-64. London: Routledge, 2020.

Duff, R. A. "Responsibility and Reciprocity." *Ethical Theory and Moral Practice* 21, no. 4 (2018/08/01 2018): 775-87.

Dworkin, Ronald. "Legal Research." *Daedalus (Cambridge, Mass.)* 102, no. 2 (1973): 53-64.

Dyson, Matthew. *Unravelling Tort and Crime.* Cambridge: Cambridge: Cambridge University Press, 2014. doi:10.1017/CBO9781107588820.

Elias, Pierre, Ash Damle, Michael Casale, Kim Branson, Chaitanya Churi, Ravi Komatireddy, and Jamison Feramisco. "A Web-Based Tool for Patient Triage in Emergency Department Settings: Validation Using the Emergency Severity Index." *JMIR Med Inform* 3, no. 2 (2015): e23-e23.

"European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence." In *2020/2014(INL)*, edited by European Parliament, 2020.

Farrell, Daniel M. "Using Wrongdoers Rightly: Tadros on the Justification of General Deterrence." *Criminal law and philosophy* 9, no. 1 (2015): 1-20.

Farthing, Sophie, John Howell, Katerina Lecchi, Zoe Paleologos, Phoebe Saintilan, and Edward Santow. "Human Rights and Technology." Camberra: Australian Human Rights Commission, 2021.

Fisher, Michael, Rafael C. Cardoso, Emily C. Collins, Christopher Dadswell, Louise A. Dennis, Clare Dixon, Marie Farrell, *et al.* "An Overview of Verification and Validation Challenges for Inspection Robots." *Robotics (Basel)* 10, no. 2 (2021): 67.

Fjelland, Ragnar. "Why General Artificial Intelligence Will Not Be Realized." *Humanities and Social Sciences Communications* 7, no. 1 (2020): 1-9.

Fletcher, George P. "Justice and Fairness in the Protection of Crime Victims." *Lewis & Clark law review* 9, no. 3 (2005): 547.

Fosch-Villaronga, Eduard, Pranav Khanna, Hadassah Drukarch, and Bart H. M. Custers. "A Human in the Loop in Surgery Automation." *Nature machine intelligence* 3, no. 5 (2021): 368-69.

French, Peter. *The Virtues of Vengeance.* University Press of Kansas, 2001.

Gasparetto, Alessandro, and Lorenzo Scalera. "From the Unimate to the Delta Robot: The Early Decades of Industrial Robotics." 284-95. Cham: Cham: Springer International Publishing, 2018.

Gless, Sabine. "Feasibility Study on a Future Council of Europe Instrument on Artificial Intelligence and Criminal Law." edited by Working Group on AI and Criminal Law - European Committee on Crime Problems (CPDC). Strasbourg, 2020.

———. "Mein Auto Fuhr Zu Schnell, Nicht Ich!“ – Strafrechtliche Verantwortung Für Hochautomatisiertes Fahren." 225-52: Nomos Verlagsgesellschaft mbH & Co. KG, 2016.

Gless, Sabine, Emily Silverman, and Thomas Weigend. "If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability." *SSRN Electronic Journal* (01/01 2016).

Graaf, Maartje de, and Bertram Malle. "How People Explain Action (and Ais Should Too)." In *AAAI Fall Symposium Series 2017*. Palo Alto, California: AAAI Press, 2017.

Greene, Erich Justin. "Effects of Disagreements between Legal Codes and Lay Intuitions on Respect for the Law." *Sci. & Engineering* 2-B, no. 64 (2003).

Gurney, Jeffrey. "Driving into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles." *Wake Forest J.L. & Pol'y* 393 (2015).

Hallevy, Gabriel. "The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Elgal Social Control." *Akron intellectual property journal* 4, no. 2 (2010): 171.

———. ""I, Robot - I, Criminal" - When Science Fiction Becomes Reality: Legal Liability of Ai Robots Committing Criminal Offenses." *Syracuse science & technology law reporter* 22 (2010): 1.

———. *Liability for Crimes Involving Artificial Intelligence Systems*. Cham: Springer International Publishing : Imprint: Springer, 2015.

———. *Liability for Crimes Involving Artificial Intelligence Systems.* 1st ed. 2015. ed. Cham: Springer International Publishing : Imprint: Springer, 2015.

———. *When Robots Kill: Artificial Intelligence under Criminal Law.* Boston: Northeastern University Press, 2013.

Hart, H. L. A. *Punishment and Responsibility: Essays in the Philosophy of Law.* 2nd ed. ed. Oxford: Oxford University Press, 2008.

Hart, H. L. A., and Tony Honoré. *Causation in the Law*. 2nd ed. ed. Oxford: Clarendon, 1985.

Hayward, K. J., and M. M. Maas. "Artificial Intelligence and Crime: A Primer for Criminologists." *Crime, Media, Culture* (2020).

Hayward, Keith J., and Matthijs M. Maas. "Artificial Intelligence and Crime: A Primer for Criminologists." *Crime, media, culture* 17, no. 2 (2021): 209-33.

Heller, Kevin Jon, and Markus Dirk Dubber. *The Handbook of Comparative Criminal Law*. Stanford, Calif.: Stanford Law Books, 2011.

Heri, Corina. "Shaping Coercive Obligations through Vulnerability: The Example of the Ecthr." Chap. 5 In *Coercive Human Rights: Positive Duties to Mobilise the Criminal Law under the Echr*, edited by Laurens Lavrysen Natasa Mavronicola. Hart Studies in Security and Justice, 93-116. Oxford: Hart Publishing, 2020.

Heyns, Christof. "Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions." UNGA, 2013.

Hildebrandt, Mireille. "Ambient Intelligence, Criminal Liability and Democracy." *Criminal law and philosophy* 2, no. 2 (2008): 163-80.

Holden, Ronald R. "Face Validity." In *The Corsini encyclopedia of psychology : Vol. 2 : D-L*, edited by Irving B. Weiner & W. Edward Craighead, 637. Hoboken, N.J: Wiley, 2010.

Holmes, Aaron. "Police Robots Keep Malfunctioning, with Mishaps Ranging from Running over a Toddler's Foot to Ignoring People in Distress." *Business Insider* (2020). doi:https://www.businessinsider.com/police-robots-security-malfunctioning-fails-knightscope-2020-1?r=US&IR=T.

Horder, Jeremy. "Gross Negligence and Criminal Culpability." *The University of Toronto law journal* 47, no. 4 (1997): 495-521.

Hruschka, Joachim. "Imputation." *Brigham Young University law review* 1986, no. 3 (1986): 669.

Hsu, Feng-Hsiung. "Ibm's Deep Blue Chess Grandmaster Chips." *IEEE MICRO* 19, no. 2 (1999): 70-81.

Hu, Ying. "Robot Criminals." *University of Michigan journal of law reform* 52, no. 2 (2019): 487.

Hunt, Elle. "Tay, Microsoft's Ai Chatbot, Gets a Crash Course in Racism from Twitter." *The Guardian* (2016). doi:https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

Husak, Douglas. "Does Criminal Liability Require an Act?". Oxford: Oxford: Oxford University Press, 2010.

Hutchinson, Terry. "Doctrinal Research: Researching the Jury." 8-39: Routledge, 2018.

Hutchinson, Terry, and Nigel Duncan. "Defining and Describing What We Do : Doctrinal Legal Research." *Deakin law review* 17, no. 1 (2012): 83-119.

Jansen, Philip, and Philip Brey. "Ethical Analysis of Rai and Robotics Technologies." In *SIENNA project - Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact*: SIENNA project, 2019.

Kamber, Kresimir. *Prosecuting Human Rights Offences: Rethinking the Sword Function of Human Rights Law*. International Criminal Law Series. Vol. 11, Leiden: Leiden: BRILL, 2017.

Karnow, Curtis E. A. "Liability for Distributed Artificial Intelligences." *Berkeley technology law journal* 11, no. 1 (1996): 147-204.

Katz, Leo, and Alvaro Sandroni. "Strict Liability and the Paradoxes of Proportionality." *Criminal law and philosophy* 12, no. 3 (2018): 365-73.

King, Thomas C., Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi. "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions." *Sci Eng Ethics* 26, no. 1 (2020): 89-120.

Kleinfeld, Joshua. "Reconstructivism: The Place of Criminal Law in Ethical Life." *Harvard law review* 129, no. 6 (2016): 1516.

"Knightscope Issues Field Incident Report." *Business Wire* (2016). doi:https://www.businesswire.com/news/home/20160713006532/en/Knightscope-Issues-Field-Incident-Report.

Köhler, Sebastian, Neil Roughley, and Hanno Sauer. "Technologically Blurred Accountability? Technology, Responsibility Gaps and the Robustness of Our Everyday Conceptual Scheme." In *Moral Agency and the Politics of Responsibility*, edited by Cornelia Ulbert, Peter Finkenbusch, Elena Sondermann and Tobias Debiel, 51-68. London: Routledge, 2017.

Koops, E. J., A. Di Carlo, L. Nocco, V. Cassamassima, and E. Stradella. "Robotic Technologies and Fundamental Rights: Robotics Challenging the European Constitutional Framework." *International journal of technoethics* 4, no. 2 (2013): 15-35.

Kowert, Weston. "The Foreseeability of Human Artificial Intelligence Interactions." *Texas law review* 96, no. 1 (2017): 181-204.

Lagioia, Francesca, and Giovanni Sartor. "Artificial Intelligence Systems under Criminal Law: A Legal Analysis and a Regulatory Perspective." *Philosophy & technology* 33, no. 3 (2020): 433-65.

Lazarus, Liora. "Positive Obligations and Criminal Justice: Duties to Protect or Coerce?". Oxford: Oxford: Oxford University Press, 2012.

Lee, Peter. "Learning from Tay's Introduction." (2016). doi:https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

Lemmens, Paul, and Marie Courtoy. "Positive Obligations and Coercion: Deterrence as a Key Factor in the European Court of Human Rights' Case Law." Chap. 3 In *Coercive Human Rights Positive Duties to Mobilise the Criminal Law under the Echr*, edited by Laurens Lavrysen and Natasa Mavronicola. Hart Studies in Security and Justice, 55-70. Oxford: Hart Publishing, 2020.

Leroux, Christophe, and Roberto Labruto. *A Green Paper on Legal Issues in Robotics*. 2012.

Leslie, David , Christopher Burr, Mhairi Aitken, Josh Cowls, Mike Katell, and Morgan Briggs. "Artificial Intelligence, Human Rights, Democracy, and the Rule of Law: A Primer." Council of Europe, 2021.

Lima, Gabriel, Meeyoung Cha, Chihyung Jeon, and Kyungsin Park. "The Punishment Gap: The Infeasible Public Attribution of Punishment to Ai and Robots." (2020).

Lima, Gabriel, Chihyung Jeon, Meeyoung Cha, and Kyungsin Park. "Will Punishing Robots Become Imperative in the Future?" In *Conference on Human Factors in Computing Systems*, 1-8: ACM, 2020.

Liu, Hin-Yan, Matthijs Maas, John Danaher, Luisa Scarcella, Michaela Lexer, and Leonard Van Rompaey. "Artificial Intelligence and Legal Disruption: A New Model for Analysis." *Law, innovation and technology* 12, no. 2 (2020): 205-58.

Liu, Hin-Yan, and Karolina Zawieska. "From Responsible Robotics Towards a Human Rights Regime Oriented to the Challenges of Robotics and Artificial Intelligence." *Ethics and information technology* 22, no. 4 (2017): 1-13.

Livermore, Joseph M., and Paul E. Meehl. "Virtues of M'naghten." *Minnesota law review* 51 (1966): 789.

"Losing Humanity: The Case against Killer Robots." Human Rights Watch & International Human Rights Clinic of the Human Rights Program at Harvard Law School, 2012.

Maas, Matthijs. "Artificial Intelligence Governance under Change: Foundations, Facets, Frameworks." 2021.

Marshall, Aarian. "Uber's Self-Driving Car Saw the Woman It Killed, Report Says." *Wired* (2018). doi:https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb-report/.

———. "Why Wasn't Uber Charged in a Fatal Self-Driving Car Crash?" *Wired*, 2020.

Matthias, Andreas. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and information technology* 6, no. 3 (2004): 175-83.

Mavronicola, Natasa. "Coercive Overreach, Dilution and Diversion: Potential Dangers of Aligning Human Rights Protection with Criminal Law (Enforcement)." Chap. 9 In *Coercive Human Rights Positive Duties to Mobilise the Criminal Law under the Echr*, edited by Laurens Lavrysen and Natasa Mavronicola. Hart Studies in Security and Justice, 183-202. Oxford: Hart Publishing, 2020.

———. "Taking Life and Liberty Seriously: Reconsidering Criminal Liability under Article 2 of the Echr." *Modern law review* 80, no. 6 (2017): 1026-51.

McAllister, Amanda. "Stranger Than Science Fiction: The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention against Torture." *Minnesota law review* 101, no. 6 (2017): 2527-73.

McQuillan, Dan. "People’s Councils for Ethical Machine Learning." *Social media + society* 4, no. 2 (2018): 205630511876830.

Metz, Cade. "In Two Moves, Alphago and Lee Sedol Redefined the Future." *Wired* (2016). doi:https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/.

Millar, Jason, and Ian Kerr. "Delegation, Relinquishment, and Responsibility: The Prospect of Expert Robots." 102-27, 2016.

"Mind the Gap: The Lack of Accountability for Killer Robots." HRW and IHRC, 2015.

Moran, Michael. "Sex Robots 'Could Be Guilty of Rape If Their Programming Fails." (2019). doi:https://www.dailystar.co.uk/news/weird-news/sex-robot-rape-international-congress-18672803

Mulligan, Christina. "Revenge against Robots." *South Carolina law review* 69, no. 3 (2018): 579.

Nissenbaum, Helen. "Accountability in a Computerized Society." *Science and engineering ethics* 2, no. 1 (1996): 25-42.

O'Sullivan, Shane, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo, Katharina Holzinger, *et al.* "Legal, Regulatory, and Ethical Frameworks for Development of Standards in Artificial Intelligence (Ai) and Autonomous Robotic Surgery." *Int J Med Robot* 15, no. 1 (2019): e1968-n/a.

Ohm, Paul. "The Fourth Amendment in a World without Privacy." *Mississippi law journal* 81, no. 5 (2012): 1309.

Osmani, N. "The Complexity of Criminal Liability of Ai Systems." *Masaryk University Journal of Law and Technology* 14, no. 1 (2020): 53-82.

Pagallo, Ugo. "Crimes." In *The Laws of Robots: Crimes, Contracts, and Torts*, 45-78. Dordrecht: Springer Netherlands, 2013.

———. *The Laws of Robots: Crimes, Contracts, and Torts.* Law, Governance and Technology Series. Vol. 10, Dordrecht: Dordrecht: Springer Netherlands, 2013.

Pagallo, Ugo, and Serena Quattrocolo. "The Impact of Ai on Criminal Law, and Its Twofold Procedures." 385-409, 2018.

Panebianco, Giuseppina. "The Nulla Poena Sine Culpa Principle in European Courts Case Law: The Perspective of the Italian Criminal Law." 47-78. Cham: Cham: Springer International Publishing, 2014.

Pasquale, Frank. "Restoring Transparency to Automated Authority." *Journal on telecommunications & high technology law* 9, no. 1 (2011): 235.

Paternoster, Raymond. "How Much Do We Really Know About Criminal Deterrence?". *Journal of Criminal Law and Criminology* 100, no. 3 (2010): 765-824.

"Presidency Conclusions - the Charter of Fundamental Rights in the Context of Artificial Intelligence and Digital Change." edited by Presidency of the Council of Europe. Brussels, 2020.

Price, W. Nicholson. "Medical Malpractice and Black-Box Medicine." 295-306: Cambridge University Press, 2018.

Rachum-Twaig, Omni. "Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots." *University of Illinois law review* 2020, no. 4 (2020): 1141-75.

Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, *et al.* "Machine Behaviour." *Nature* 568, no. 7753 (2019): 477-86.

"Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics." In *2015/2103(INL)*, edited by European Parliament, 2017.

Reyes Romero, Ítalo. "Un Concepto De Riesgo Permitido Alejado De La Imputación Objetiva." *Ius et Praxis* 21, no. 1 (2015): 137-69.

Rinie van Est, Joost Gerritsen. "Human Rights in the Robot Age: Challenges Arising from the Use of Robotics, Artificial Intelligence, and Virtual and Augmented Reality – Expert Report Written for the Committee on Culture, Science, Education and Media of the Parliamentary Assembly of the Council of Europe." The Hague: Rathenau Instituut, 2017.

Robinson, Darryl. "The Identity Crisis of International Criminal Law." *Leiden Journal of International Law* 21, no. 4 (2008): 925-63.

Robinson, Paul H. *Distributive Principles of Criminal Law : Who Should Be Punished, How Much?* Oxford: Oxford University Press, 2008.

———. *Intuitions of Justice and the Utility of Desert.* New York: New York: Oxford University Press, 2013. doi:10.1093/acprof:oso/9780199917723.001.0001.

Robinson, Paul H., and John M. Darley. "Intuitions of Justice: Implications for Criminal Law and Justice Policy." *Southern California law review* 81, no. 1 (2007): 1-67.

Rodrigues, Rowena. "Legal and Human Rights Issues of Ai: Gaps, Challenges and Vulnerabilities." *Journal of Responsible Technology* 4 (2020).

Roxin, Claus. *Strafrecht : Allgemeiner Teil : 1 : Grundlagen. Der Aufbau Der Verbrechenslehre.* 2. Aufl. ed. Vol. 1, München: Beck, 1994.

Royakkers, L. M. M., and van Q. C. Est. "A Literature Review on New Robotics : Automation from Love to War." *International Journal of Social Robotics* 7, no. 5 (2015): 549-70.

Salako, Solomon E. "Strict Criminal Liability: A Violation of the Convention?". *Journal of criminal law (Hertford)* 70, no. 6 (2006): 531-49.

Santoni De Sio, F. "Killing by Autonomous Vehicles and the Legal Doctrine of Necessity." *Ethical theory and moral practice* 20, no. 2 (2017): 411-29.

Santoni de Sio, Filippo, and Giulio Mecacci. "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them." *Philosophy & technology* (2021).

Sathykumar, Kaviya, Michael Munoz, Jaikaran Singh, Nowair Hussain, and Benson A. Babu. "Automated Lung Cancer Detection Using Artificial Intelligence (Ai) Deep Convolutional Neural Networks: A Narrative Literature Review." *Curēus (Palo Alto, CA)* 12, no. 8 (2020): e10017-e17.

Schrempf, M., and M. Anthuber. "Autonomous Surgery—a Vision of the Future." *Chirurg* 90, no. 11 (2019): 937.

Seibert-Fohr, Anja. *Prosecuting Serious Human Rights Violations.* Oxford: Oxford University Press, 2009.

Selbst, Andrew D. "Negligence and Ai's Human Users." *Boston University law review* 100, no. 4 (2020): 1315-76.

Selbst, Andrew D., and Solon Barocas. "The Intuitive Appeal of Explainable Machines." *Fordham law review* 87, no. 3 (2018): 1085-139.

"Shaking the Foundations: The Human Rights Implications of Killer Robots." HRW and IHRC, 2014.

Sharkey, Noel, Marc Goodman, and Nick Ross. "The Coming Robot Crime Wave." *Computer (Long Beach, Calif.)* 43, no. 8 (2010): 116-15.

Simester, A. P. "Is Strict Liability Always Wrong?". Oxford: Oxford: Oxford University Press, 2005.

Simmler, Monika, and Nora Markwalder. "Guilty Robots? – Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence." *Criminal law forum* 30, no. 1 (2018): 1-31.

Simpson, Richard, Edmund LoPresti, Steve Hayashi, Illah Nourbakhsh, and David Miller. "The Smart Wheelchair Component System." *J Rehabil Res Dev* 41, no. 3 B (2004): 429-42.

Spencer, J. R., and Marie-Aimée Brajeux. "Criminal Liability for Negligence—a Lesson from across the Channel?". *ICLQ* 59, no. 1 (2010): 1-24.

Stark, Findlay. *Culpable Carelessness: Recklessness and Negligence in the Criminal Law*. Cambridge: Cambridge University Press, 2016.

"Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control." Human Rights Watch, 2020.

Stoyanova, Vladislava. "Article 4 of the Echr and the Obligation of Criminalising Slavery, Servitude, Forced Labour and Human Trafficking." *Cambridge International Law Journal* 3, no. 2 (2014): 407-43.

Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning : An Introduction*. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press, 1998.

Swart, Bert, and Antonio Cassese. "Part a Major Problems of International Criminal Justice, Ii Fundamentals of International Criminal Law, Modes of International Criminal Liability." Oxford University Press, 2009.

Tadros, Victor. *Criminal Responsibility*. Oxford: Oxford University Press, 2007.

"Tesla: Elon Musk Suggests Autopilot Not to Blame for Fatal Crash." *BBC* (2021). doi:https://www.bbc.com/news/technology-56799749.

Thomasen, Kristen. "Examining the Constitutionality of Robot-Enhanced Interrogation." In *Robot Law*, edited by Ryan Calo, Michael Froomkin and Ian Kerr, 306-30. Cheltenham, England: Edward Elgar Publishing, 2016.

Tyler, Tom R. *Why People Obey the Law*. New Haven: Yale University Press, 1990.

"Uber's Self-Driving Operator Charged over Fatal Crash." *BBC* (2020). doi:https://www.bbc.com/news/technology-54175359.

UNGA. "Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law." New York, 2005.

van der Wilt, H. "Nullum Crimen and International Criminal Law: The Relevance of the Foreseeability Test." *Nordic journal of international law = Acta scandinavica juris gentium* 84, no. 3 (2015): 515-31.

Van Hoecke, Mark. "Legal Doctrine: Which Method(S) for What Kind of Discipline?". In *Methodologies of Legal Research : What Kind of Method for What Kind of Discipline?*, edited by Mark Van Hoecke, 1-18. London: Hart Publishing, 2011.

van Klink, B. M. J., and H. S. Taekema. "On the Border: Limits and Possibilities of Interdisciplinary Research." 7-32: Mohr Siebeck, 2011.

Vanacore, Giulio. "Legality, Culpability and Dogmatik: A Dialogue between the Ecthr, Comparative and International Criminal Law." *International criminal law review* 15, no. 5 (2015): 823-60.

Velasquez, Manuel. "Debunking Corporate Moral Responsibility." *Business ethics quarterly* 13,47, no. 4 (2003): 155,531,294,471-562,299,172,481.

Vincent, James. "Have Autonomous Robots Started Killing in War?" *The Verge* (2021). doi:https://www.theverge.com/2021/6/3/22462840/killer-robot-autonomous-drone-attack-libya-un-report-context.

———. "Mall Security Bot Knocks Down Toddler, Breaks Asimov's First Law of Robotics." *The Verge* (2016). doi:https://www.theverge.com/2016/7/13/12170640/mall-security-robot-k5-knocks-down-toddler.

———. "Twitter Taught Microsoft's Ai Chatbot to Be a Racist Asshole in Less Than a Day." *The Verge* (2016). doi:https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Watson, Gary. "Two Faces of Responsibility." *Philosophical Topics* 24, no. 2 (1996): 227-48.

Waxman, Kenneth Anderson and Matthew C. "Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can." In *Jean Perkins Task Force on National Security and Law Essay Series*, edited by The Hoover Institution Stanford University, 2013.

Weyns, Danny, Elke Steegmans, and T. O. M. Holvoet. "Towards Active Perception in Situated Multi-Agent Systems." *Applied artificial intelligence* 18, no. 9-10 (2004): 867-83.

Winfield, Alan F. T., Serena Booth, Louise A. Dennis, Takashi Egawa, Helen Hastie, Naomi Jacobs, Roderick I. Muttram, *et al.* "Ieee P7001: A Proposed Standard on Transparency." *Frontiers in robotics and AI* 8 (2021): 665729-29.

Yampolskiy, Roman. "Unpredictability of Ai." *Journal of Artificial Intelligence and Consciousness* 7 (2019): 109-18.

Yeung, Karen. "A Study of the Implications of Advanced Digital Technologies (Including Ai Systems) for the Concept of Responsibility within a Human Rights Framework." Council of Europe, 2010.

Zimmer, Alf C. "Can Autonomous Cars Improve the Safety and Efficiency in Road Traffic?". *Automatisierungstechnik : AT* 65, no. 7 (2017): 458-64.

## B. Laws and technical guidelines

### *i. National standards*

Law no. 2019-486 of 22 May 2019 on the Growth and Transformation of Companies, Art. 125 modifying ordinance No. 2016-1057 on the Testing of Vehicles with Delegation of Driving in Public Roads.

### *ii. Technical standards and ethical guidelines for AI development and use*

Draft Standard for Transparency of Autonomous Systems, IEEE P7001 (IEEE, PAR approval: 2016-12-07).

Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems, BS 8611:2016 (BSI, first published April 2016, currently under review).

### *iii. International instruments*

Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, 4 November 1950, ETS 5.

International Committee of the Red Cross (ICRC), Geneva Convention Relative to the Protection of Civilian Persons in Time of War (Fourth Geneva Convention), 12 August 1949, United Nations, Treaty Series, vol 287, p. 75.

Organization of African Unity (OAU), African Charter on Human and Peoples' Rights ("Banjul Charter"), 27 June 1981, CAB/LEG/67/3 rev. 5, 21 I.L.M. 58 (1982).

Organization of American States. 1969. “American Convention on Human Rights.” *Treaty Series, No. 36*. San Jose: Organization of American States.

UN General Assembly, Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment, 10 December 1984, United Nations, Treaty Series, vol. 1465, p. 85.

UN General Assembly, Convention on the Prevention and Punishment of the Crime of Genocide, 9 December 1948, United Nations, Treaty Series, vol. 277, p. 78.

UN General Assembly, Convention on the Rights of Persons with Disabilities: resolution / adopted by the General Assembly, 24 January 2007, A/RES/61/106.

UN General Assembly, International Covenant on Civil and Political Rights, 16 December 1966, United Nations, Treaty Series, vol. 999, p. 171.

### A. Judicial decisions and views of treaty bodies

#### *i. African Commission on Human and People's Rights*

*Egyptian Initiative for Personal Rights & Interights v. Egypt (Communication), no. 323/06, 2011.*

*Kevin Mgwanga Gunme and others., v. Cameroon* (Communication), no. 266/03, 2009.

*Monim Elgak, Osman Hummeida and Amir Suliman v. Sudan (Communication), no. 379/09, 2014.*

*Mouvement Burkinabé des Droits de l'Homme et des Peuples v Burkina Faso* (Communication), no. 204/97, 2001.

*Noah Kazingachire, John Chitsenga, Elias Chemvura andBatanai Hadzisi v. Zimbabwe* (Communication), no. 295/04, 2012.

*The Social and Economic Rights Action Center and the Center for Economic and Social Rights v. Nigeria*, (Communication), no. 155/96, 2001.

#### *ii. African Court on Human and People's Rights*

*Beneficiaries of the Late Norbert Zongo and others v. Burkina Faso,* no. 013/2011, 2014.

#### *iii. European Court of Human Rights*

*A v. The United Kingdom, no. 25599/94, ECHR, 1998-VI.*

*Ali and Ayşe Duran v. Turkey,* no. 42942/02, ECHR, 2008

*Asiye Genç v. Turkey*, no. 24109/07, ECHR, 2015.

*Beganović v. Croatia*, no. 46423/06, ECHR, 2009.

*Calvelli and Ciglio v. Italy, no. 32967/96, ECHR, 2002-I.*

*Cantoni v. France*, no. 17862/91, ECHR, 1996-V.

*Del Río Prada v. Spain [GC],* no. 42750/09, ECHR, 2013.

*Engel and Others v. the Netherlands,* no. 5100/71; 5101/71; 5102/71; 5354/72; 5370/728, ECHR, 1976, Series A no. 22.

*Estamirov and others v. Russia*, no. 60272/00, ECHR, 2006.

*G. v. France,* no. 15312/89, ECHR, A325-B.

*GIEM SRL and Others v. Italy* [GC], nos. 1828/06 and 2 others, ECHR, 2018

*Jussila v. Finland* [GC], no. 73053/01, ECHR 2006-XIV.

*Kononov v. Latvia* [GC], no. 36376/04, ECHR, 2010.

*KU v. Finland*, no. 2872/02, ECHR, 2008.

*MC v Bulgaria*, no. 39272/98, ECHR, 2003-XII (extracts).

*Oruk v. Turkey*, no. 33647/04, ECHR, 2014.

*Osman v. The United Kingdom* [GC], no. 23452/94, ECHR, 1998-VIII.

*Öneryildiz v. Turkey*, no. 48939/99, ECHR, 2004-XII.

*Sabalić v. Croatia*, no. 50231/13, ECHR, 2021.

Salabiaku v. France, no. 10519/83, § 26, ECHR, 1988 A141-A.

*Sinim v. Turkey*, no. 9441/10, ECHR, 2017.

*Smiljanić v. Croatia*, no. 35983/14, ECHR, 2021.

*Söderman v. Sweden*, no. 5786/08, ECHR, 2013

*Volodina v. Russia,* no. 41261/17, ECHR, 2019.

*X & Y v. The Netherlands*, no. 8978/80, ECHR, A91.

*Yasa v Turkey*, no. 22495/93, ECHR, 1999-VI.

*iv. Human Rights Committee (views)*

*Marija and Dragana Novaković v. Serbia,* no. 1556/2007, CCPR/C/100/D/1556/2007, 2010.

*Kazantzis v. Cyprus*, no. 972/2001, CCPR/C/78/D/972/2001, 2003.

*Rajapakse v. Sri Lanka*, no. 1250/2004, CCPR/C/87/D/1250/2004, 2006.

*Paksas v Lithuania*, no. 2155/2012, CCPR/C/110/D/2155/2012, 2014.

*v. Inter-American Court of Human Rights*

*Albán-Cornejo et al. v. Ecuador*, no. 12406, 2007.

*Velásquez Rodríguez v. Honduras,* no. 7920, 1988.