Moral responsibility for unforeseen harms caused by autonomous systems

by

Zoë Larissa Mayne Porter

Department of Philosophy, University of York

Thesis submitted for the degree of Doctor of Philosophy

June 2021

Abstract

Autonomous systems are machines which embody Artificial Intelligence and Machine Learning and which take actions in the world, independently of direct human control. Their deployment raises a pressing question, which I call the *locus of moral responsibility* question: who, if anyone, is morally responsible for a harm caused directly by an autonomous system? My specific focus is moral responsibility for *unforeseen* harms.

First, I set up the *locus of moral responsibility* problem. Unforeseen harms from autonomous systems create a problem for what I call the Standard View, rooted in common sense, that human agents are morally responsible. Unforeseen harms give credence to the main claim of 'responsibility gap' arguments – that humans do not meet the control and knowledge conditions of responsibility sufficiently to warrant such an ascription.

Second, I argue a delegation framework offers a powerful route for answering the *locus of moral responsibility* question. I argue that responsibility as *attributability* traces to the human principals who made the decision to delegate to the system, notwithstanding a later suspension of control and knowledge. These principals would also be *blameworthy* if their decision to delegate did not serve a purpose that morally justified the subsequent risk-imposition in the first place. Because I argue that different human principals share moral responsibility, I defend a pluralist Standard View.

Third, I argue that, while today's autonomous systems do not meet the agential condition for moral responsibility, it is neither conceptually incoherent nor physically impossible that they *might*. Because I take it to be a contingent and not a necessary truth that human principals exclusively bear moral responsibility, I defend a soft, pluralist Standard View.

Finally, I develop and sharpen my account in response to possible objections, and I explore its wider implications.

Acknowledgements

For their support and wise counsel, I thank my supervisors, Professor Stephen Holland and Professor Paul Noordhof. I also thank Dr. Janet Eldred and Julie Kay for their kindnesses over the course of my PhD. Thanks are due also to Ivan Kyambadde, for sharing the teaching and marking load with me on the 'Readings in the Ethics of AI' module in the Summer Term before this thesis was submitted.

I owe a huge debt of gratitude to Professor John McDermid and Dr. Ibrahim Habli, for their encouragement and many stimulating and helpful conversations - as well as for relieving me of the duties of my role as a Research Associate in the Assuring Autonomy International Programme for the six weeks prior to submission. I could not hope for better colleagues.

Words cannot convey the heroic levels of support my family, and particularly my husband, have given me. Thank you all for your patience, and your love, and for letting me get on with it. Thanks also to my friends, particularly Adam and Celia.

For my parents

Author's declaration

I declare that this thesis is a presentation of original work and I am the sole author. The work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Early versions of some of the material in this thesis have been published in the following places:

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., and Porter, Z., 2020. Mind the gaps: assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence*, 279, 103201

Habli, I., Lawton, T. & Porter, Z., 2020. Artificial intelligence in health care: accountability and safety. *Bulletin of the World Health Organization*, 98(4), pp. 251-256

Porter, Z., Habli, I., Monkhouse, H. & Bragg, J., 2018. The moral responsibility gap and the increasing autonomy of systems. *International Conference on Computer Safety, Reliability, and Security*, pp. 487-493

This thesis marks a development of my position in those papers.

CONTENTS

INTRODUCTION

'Autonomous system' is a term of art in engineering. Autonomous systems are machines which embody Artificial Intelligence (AI) and especially Machine Learning (ML) and which take actions in the world, independently of direct human control. Developed and deployed for increasingly critical tasks, autonomous systems can have a direct impact on human life, liberty, and wellbeing.

This raises a question: who is morally responsible when an autonomous system directly causes a harm to a person, and on what grounds? I call this the *locus of moral responsibility* question. Specifically, my concern is with responsibility for harms that were unforeseen by the relevant human agents.

To set the scene, imagine the following hypothetical scenarios.

- One bright, spring day Jacinta is travelling in her self-driving car to the supermarket. She has grown fond of the car and calls it Myrtle. Operating in fully autonomous mode, the car avoids a traffic jam on the main route and takes a detour down a street by a park. The street is lined with trees, casting striped shadows across the road. A small child runs out into the road from behind some parked cars. Due to the intermittent sunlight, the car's front sensors do not detect the child immediately. Once they do, a few seconds later, the vehicle issues a transition demand to Jacinta. But she has fewer seconds than she needs to take over control of the vehicle, and the self-driving car hits and injures the small child. One distraught mother, the child's sibling, and a pushchair laden down with shopping stands on the roadside.

- A military superpower has invested heavily in missiles with ML-enabled capability. Its Army has deployed one of these missiles to locate and destroy the munitions factory of a terrorist cell in a remote mountainous region. The autonomous system directly receives, as structured data, human intelligence and cyber-intelligence about the target. It locates the factory, and immediately destroys it – monitored by a military command centre in a nearby region. No one on the ground is killed. But debris from the explosion blocks and contaminates the only source of clean drinking water for a village at the foothills of the

mountain. One hundred people's lives are gravely endangered until emergency aid is brought to them.

- Morgan, a stellar high school student with dreams of academia, has become fond of using an advanced robotic personal assistant – the 'Scholar Bot' – when writing essays. Users interact with the bot verbally, like a spoken search engine, and it adapts to user preferences over time. The bot has access to all of the main academic resources, as well as the full 20-volume Oxford English Dictionary. It can download source material to the user's laptop and insert citations directly into documents. It can also send journal articles and book chapters to the printer on a voice command. Morgan is in scholar's heaven. The bot's developers have taken pains to ensure that this application is not connected to any media from which racist or sexist language might emerge. Our high school student is writing an advanced-level essay on the Holocaust. After a useful overview of the Nuremberg Trials, the Scholar Bot starts to provide information on the Nuremberg Code, created in the wake of the discovery of Nazi human experimentation in concentration camps. It then provides, in increasingly graphic detail, accounts of the worst of these medical experiments conducted on children. It prints a couple of images, on a regretted voice command. Morgan, who is only seventeen, is traumatised by this graphic and extreme content and suffers flashbacks for many weeks.

- In a hidden corner of the North-East of England is one of the best care homes in the country. The care home manager, David, has been integrating technology successfully into the home for several years, particularly for physical assistance tasks, which both helps the residents and relieves staff of heavy and exhausting work. The home has now started to deploy assistive lifting robots for its elderly residents in their rooms, to stand up from a sitting position. The robots have been developed in collaboration with gerontologists and physical therapists to ensure that they are safe and suitable for use. The robots are adaptive. Once in an environment, they adapt pressure and speed according to user preferences. But on one occasion, the recipient, Tom, expresses his preferences in an unusual way, and, for no discernible reason, this, when combined with the loud noise from a lawnmower in the garden outside, causes the robot to apply excessive pressure and speed, wrenching Tom's shoulders and wrists. He is left in severe pain.

What should we say about who bears moral responsibility in these cases? Most of us will share the intuition that it must be human agents who are morally responsible in any cases involving technological systems. I call this the Standard View. It is less straightforward to explain, on principled grounds, with reference to autonomous systems *which* humans are morally responsible and *why*, or on what *grounds*. That is what the *locus of moral responsibility* question seeks to establish.

The question is a practically pressing one. Autonomous systems are deployed or being developed for deployment in many safety-critical and security-critical domains. Frameworks for tracing moral responsibility for any harms these systems may cause, yielding non-arbitrary ascriptions of moral responsibility to the appropriate human agents, are needed as a matter of justice. Sometimes the harm will have been non-negligently unforeseen by the relevant human agents, as is arguably the case in the four hypothetical examples just given. In unravelling the answer to *the locus of moral responsibility* question in these hard cases, we may also provide clarity for conscientious human agents who are unsure what, if anything, it would be fair to blame them for when things go wrong, when they have been rigorous and diligent in their decision-making about the system.

The question is philosophically interesting, too. It touches on the conceptual liminality of autonomous systems. As Johnson & Verdicchio put it, "It is not that we have discovered a new entity and are poking and prodding it to figure out what it is. Rather we are at once creating a new type of entity and simultaneously asking what it is" (2018, p. 292). The project of doing so is complicated by those properties of autonomous systems that seem to place them at the threshold of certain conceptual categories. For instance, they are like tools in some respects, like intelligent agents in others. They have sufficient agency to cause morally significant harm, independently of human intervention, yet they fall short of the agency required to bear any moral responsibility for that harm. And as they grow in complexity, and become adaptive and embedded in our wider moral systems, our traditional conceptual categories for reasoning about their properties and agential capacities may well evolve in unexpected ways.

My central claim is that a delegation framework is the best framework for ascribing moral responsibility for harms caused by autonomous systems, including in the tricky cases of responsibility for unforeseen harmful consequences. My argument owes a debt to the

thinking of Di Nucci (2020) and Di Nucci & Santoni de Sio (2014), whom I take to be the progenitors of the position that delegation specifically as an action-theoretic concept is a powerful device for understanding our relations with technological systems.

A defence of the Standard View – which I characterise more precisely as the position that moral responsibility is borne completely and exclusively by human agents – faces a challenge from what has come to be known as the 'responsibility gap' problem (Sparrow 2007, Matthias 2004). The worry here is that no human has sufficient control over, or awareness of, the system's processes fairly to be ascribed moral responsibility for the consequences of its outputs. Since control and knowledge are generally taken to be necessary conditions of moral responsibility, this claim, if true, presents a serious moral responsibility problem. This problem is most forceful in the case of unforeseen harms.

I construct a delegation framework to answer the *locus of moral responsibility* question in response to this challenge to the Standard View. Drawing on a distinction made by Scanlon between 'responsibility as attributability' and 'blameworthiness' (1998, p. 248), I answer the question in two parts. First, I determine the *loci of moral responsibility as attributability*. This establishes to whom the harm is attributable in a way that is necessary to provide the basis for their moral appraisal. Here, the delegation relation furnishes a tracing principle that warrants attributions of moral responsibility to the human agents who participate in the complex decision to delegate to an autonomous system. This tracing principle constitutes an exception to the necessity of the control and knowledge conditions. Second, I determine the *loci of moral responsibility as blameworthiness*. This establishes whether the principals identified by the preceding step are also blameworthy for the harm. I argue that the threshold for moral blame is low. That is, my account is morally strict. My claim here is that the principals would clearly be blameworthy if they had been negligent, but they would also be blameworthy if they had not incorporated, in addition to the weighing of the severity and likelihood of risk, moral values into the assessment of the risk imposed by the deployment of the autonomous system *and* if the decision to deploy the system does not serve a purpose that morally justifies the subsequent risk imposition.

The delegation framework provides the grounds for warranted ascriptions of moral responsibility to human agents, and thereby a defence of the Standard View – even in the hard cases of unforeseen harms. More specifically, the account I propose is a pluralist and

soft Standard View. By a 'pluralist' Standard View, I mean that moral responsibility is shared between members of different kinds or classes of human principal. In the real-world, the decision to delegate to an autonomous system is multi-levelled and complex. I argue that its complexity means that there are in fact two classes of human principal agent: the low-level principal who delegates on the ground, and the high-level principals who design, constrain and approve the autonomous system for this purpose, which is its very function, higher upstream in the development process. Responsibility as attributability traces back to both classes of human principal. By a 'soft' Standard View, I mean that the exclusivity of human moral responsibility for harms caused by autonomous systems is a contingent but not a necessary truth.

This brings me to the final part of my argument, which considers whether autonomous systems themselves could ever be morally responsible for harms they cause – and, if so, when and how. This question of machine responsibility is linked to the core argument advancing a delegation framework for ascribing moral responsibility because what is at stake is whether autonomous systems could ever share moral responsibility *qua* delegates or proxies with their human principals – just as human delegates or proxies sometimes share moral responsibility with their principals. I argue that, though current autonomous systems are not moral agents and so a precondition for their moral responsibility is not met, it is physically possible that suitably adaptive future autonomous systems, embedded in a wider moral system, may learn or evolve to behave in ways that are genuinely other-regarding or altruistic. This, I argue, would be sufficient for an ascription of minimal, marginal, moral responsibility to the systems. This is a speculative argument, but it indicates the possibility that the Standard View may one day *have* to give way to non-Standard notions of shared human and artificial moral responsibility and, as such, the argument I defend is a soft Standard View. Revealingly, even the prospect of future machine responsibility is accommodated by a delegation framework, which provides a mechanism for warranted ascriptions of moral responsibility in both cases of foreseen and unforeseen harm, and also in present day and possible future cases of autonomous systems. This normative flexibility commends it.

Having given the broad outline of my argument, I now give a Chapter-by-Chapter breakdown of the thesis.

In the **first** Chapter, I elucidate the concept of machine autonomy. I describe it as having three dimensions: independence from direct human intervention; independence from continuous human oversight; and independence from explicit human instruction. Increasing autonomy in each of these dimensions puts pressure on our traditional conceptual frameworks for ascribing moral responsibility to human agents. As such, increasing machine autonomy poses a problem for the position, deeply rooted in common sense, that humans are completely and exclusively morally responsible for harms caused by all technological systems – which I call the Standard View position. The argument I defend is situated within that position.

In the **second** Chapter, I formulate the *locus of moral responsibility* question. This is the question of who, if anyone, is morally responsible for harm caused directly by an autonomous system and on what grounds? The answer will be expressible as a three-term relation: S is morally responsible for X in virtue of P. Traditionally, these properties P, in virtue of which a responsibility ascription is warranted, include control over, and awareness of, the thing for which responsibility is sought. I clarify that X is an unforeseen harm and that my procedure for answering the *locus of moral responsibility* question will be first to discern the *locus of moral responsibility as attributability for X*, and second to determine the *locus of moral responsibility as blameworthiness for X*.

In the **third** Chapter, I set out a common objection, found in the responsibility gap literature, that the necessary properties P of the relation S bears to the consequence are not adequately met by humans to warrant ascribing moral responsibility to them for harms caused by autonomous systems. A main premise of most responsibility gap positions is that human control and knowledge are to some relevant degree suspended with respect to the behaviour of autonomous systems. They infer from this a lacuna in responsibility coverage for any harms thereby caused. This Chapter examines whether that premise is true, and concludes that, while it is not universally true, it is true in some cases, including cases of unforeseen harms.

In the **fourth** Chapter, I defend the Standard View position in the face of this challenge. While the premise that the Aristotelian conditions are suspended is sometimes true, we should not infer from this a responsibility gap. This is because fulfilment of these conditions over the immediate causal antecedent of a harm is *not* a necessary condition for warranted

ascriptions of moral responsibility to human agents in these cases. Using a delegation framework, I derive a principal-proxy tracing relation for determining the *loci of moral responsibility as attributability*, even for unforeseen harms. Within the complex, multi-levelled decision to delegate to autonomous system, both low-level principals on the ground and high-level principals upstream in the development cycle bear attributional responsibility for the consequences of the system's behaviour. I further argue that, given the nature of the risk that is committed to by the principal agents, the heuristic principle for determining the *loci of moral responsibility as blameworthiness* should be whether the decision to delegate to the system serves a purpose that morally justifies the imposition of the risk. The delegation framework I propose therefore yields a pluralist Standard View position – that a plurality of human agents are morally responsible for harms caused by autonomous systems, and that this responsibility coverage is complete.

In the **fifth** Chapter, I consider the exclusivity requirement of the Standard View position, as I have described it. This is the requirement that humans are not only completely morally responsible for harms caused by autonomous systems, but that they are exclusively so. Given that, in cases of human-to-human delegation, human delegates or proxies sometimes share moral responsibility with their principals, I consider whether this could also be true in the case of artificial proxies. I consider whether machine moral responsibility is conceptually coherent or physically possible. I argue that it is both: that some foreseeable systems could legitimately be regarded as 'marginal' morally responsible agents. As such, should autonomous systems of this kind ever come to pass and be deployed, they would not be exempt, *qua* proxies, from sharing moral responsibility with human principals for some of the harms that they cause. The exclusivity of human moral responsibility that I have endorsed is contingent on the way the world is now; it is not a necessary truth. This concludes my presentation of a soft, pluralist Standard View.

In the **sixth** Chapter, I give a summary statement of my argument, and address some key objections to it: that the delegation framework is redundant; that my individuation of the principals belies the true complexity of human relations to the system; and that my account is too demanding. Finally, I consider the further normative implications of my argument: the prospective duties it places on principals; its relation to the question of vicarious liability in law; and its conceptual implications.

To conclude, I defend a soft, pluralist Standard View position in answer to the *locus of moral responsibility* question. This position is grounded in a delegation framework for ascribing moral responsibility for harms caused by autonomous systems. Though constructed in response to a specific and thorny sub-set of harm, the unforeseen harm, the delegation framework generalises to all cases – those that are foreseen, those that unforeseen, and those involving autonomous systems with more diverse and developed agential capacities than the autonomous systems of today.

CHAPTER 1
AUTONOMOUS SYSTEMS

In the first Chapter, I define an 'autonomous system' and elucidate the concept of 'machine autonomy', which is characterised in terms of degrees of independence from direct human control. I introduce a position I call the 'Standard View'. This is the position that moral responsibility for any harms directly caused by an autonomous system would rest completely and exclusively with human agents. The argument of this thesis is situated within that position, and it defends the Standard View even in the case of unforeseen harms.

1.1 What is an autonomous system?

Autonomous systems are physical or digital machines which embody Artificial Intelligence (AI) and especially Machine Learning (ML). Crucially, these systems replace the human agent in the fulfilment of all or some of a decision-making task, and they implement their outputs, or take actions, directly in the world, independently of direct human control. The paradigmatic example is a fully self-driving car, which plans and executes a whole journey through a dynamic environment such as a city, without the operational control of a human driver. The replacement of the human in such a task is what Searle calls the system's 'agentive function'. These are "functions that we do not discover, and that do not occur naturally, but that are assigned relative to the practical interests of conscious agents" (Searle, 1995, p. 20). While Searle gives examples of more rudimentary artefacts, such as chairs and screwdrivers, the concept of agentive function extends to complex computational systems, too. But whereas the agentive function of simple tools is to be used *by* these conscious agents, the agentive function of autonomous systems is to undertake tasks *for* them, to some degree. This preposition, 'for', signals something crucial to our inquiry; it is central to the relation of delegation.

Since the start of the 21st century, the development of autonomous systems has intensified and their deployment is increasingly widespread. For the most part, today's deployed systems are not fully independent of direct human control but are assistance or advisory systems. But the direction of travel in research and development is towards systems that have greater degrees of independence from humans in-the-loop (Fischer *et al.*, 2021; Law Commission

2020). Governments and businesses are investing heavily in developing autonomous systems for a wide range of tasks, sectors, and operating environments.[1] COVID-19 has only accelerated this innovation, with national lockdowns highlighting the role that these systems can play in logistics and the delivery of vital supplies (Royal Academy of Engineering, 2020a). Many of these tasks and environments – such as in the military, or automotive, or health and social care, or criminal justice – are those in which the risks of harm or the promise of benefit from an autonomous system's outputs are morally significant, concerning the life, wellbeing, or liberty of agents with moral standing, whose needs and interests have moral weight. AI-enabled weapons and defence systems (AWS) exist, despite widespread outcry, and no internationally binding veto has been placed on such systems having a lethal agentive function (United Nations, 2014). But, equally, there are autonomous robots that clear minefields before a human has to cross them, and carry the equipment of soldiers across difficult terrains, to the gratitude of many army platoons (Carpenter, 2016). Accidents on the roads are inevitable, and self-driving cars not only sometimes may, but already have, directly killed and injured people (National Transportation Safety Board, 2019; Nyholm & Smids, 2016; Goodall, 2014). But the hope is also that the deployment of autonomous cars will radically reduce the vast number of fatalities due to erratic, drunken, or unlawful human driving. The prospect of robots in policing and prisons – as well as decision support systems and facial recognition technology to inform critical decisions about a person's treatment under the law – raise uneasy questions about the sanctity of human judgement in criminal justice (Pasquale, 2020, pp. 128-131). But softbots that classify the criminal severity of child pornography images also relieve human police officers of a psychologically damaging task. My purpose in this thesis is not to undermine the benefits that autonomous systems can bring, some of which will surely be morally compelling, either because of the purpose the systems serve, or the relief to human agents that they provide. But my focus is on the harms caused by autonomous systems. The logical object of the *locus of moral responsibility* question is a *harm* to a person, directly caused by an autonomous system, whether this is damage to her physical or psychological wellbeing or an injustice of some other kind.

The first stage of this inquiry is to provide a definition of 'autonomous system' and to elucidate the different dimensions of machine autonomy. The standard definitions of

[1] The UK Government, for example, has invested more than £2.3bn into AI across a range of initiatives since 2014. This figure excludes defence spend (HM Government, UK National AI Strategy 2021).

autonomous systems given by engineers and computer scientists involve reference to two central features: 'autonomy' and 'decision-making':

> *"A computational system is called* autonomous *if it is able to make its own decisions, or take its own actions, without human supervision or control."*
>
> (Fisher *et al.*, 2021, p. 1)

> *"Autonomous systems make informed decisions for themselves in complex environments. The systems can have a physical manifestation such as a vehicle or a robot or be purely software based, such as a financial trading algorithm or a medical decision-support system."*
>
> (Royal Academy of Engineering, 2020b, p.4)

> *"By 'autonomous system', we mean a system that makes decisions independently of human control."*
>
> (Burton *et al.*, 2020, p. 1)

According to our common philosophical conceptions, the correct ascription of psychological predicates such as 'autonomy' and 'decision-making' implies that one is a responsible agent, answerable for one's actions, and a fitting subject of the moral appraisal of others. Here, 'autonomy' is generally elliptical for 'moral autonomy', that inviolable aspect of a person's sense of self, intimately connected to her dignity, capacity for self-knowledge, critical reflection, and responsibility (Dworkin, 1988, p. 6). As Dworkin notes, its use as a term of art in philosophy to make sense of a tangled web of intuitions and claims is too broad to admit of a single definition, but the idea of autonomous persons as self-determining is widely shared (Dworkin, 1988, p. 7 & p. 9). It involves the capacity to direct one's own life, as well as the condition of doing so when the circumstances permit (Feinberg, 1989, p. 27). This capacity and condition is presented by Berlin as "positive freedom": "deciding, not being decided for, self-directed [...] conceiving goals and policies of my own and realizing them" (Berlin, 1969, p. 131).

The routine ascription of these predicates to 'autonomous systems' in engineering and computer science is not taken to have those connotations. Positive freedom contrasts with "negative freedom", which is the absence of external obstacles, constraints, or interference from others in the pursuit of one's self-directed goals and activities. Though not an exact mapping onto Berlin's distinction (because his negative freedom specifically concerns the

independence of free, self-determining *human* agents from external interference), we can make a *similar* distinction between positive and negative autonomy, and apply it to the difference between humans and autonomous systems. People have positive autonomy, and fortunate people have both positive and negative autonomy, but the autonomy of autonomous systems is, presently, purely negative. As Roff puts it:

> *"Machine autonomy is merely the ability to act in the world without someone or something immediately directing that action. There is no discussion of intent, of consciousness, or of the type of freedom that admits of moral operators of praise or blame. The robotic notion of autonomy is radically minimalist, as it removes ethical evaluation by definitional fiat."*
>
> (Roff, 2013, pp. 353-4)

In Chapter 5, I consider whether those moral operators of praise and blame will always exclude autonomous systems. But my working assumption is that the machines that exist today, and in the foreseeable future, do not meet this criterion. The scope of machine autonomy – as I elucidate it below – is restricted to negative autonomy.[2]

Before providing this elucidation of machine autonomy I would like to make explicit how I use some key terms throughout this thesis. Specifically, I tend to speak of a system's 'outputs' and 'behaviour' rather than it's 'decisions' and 'actions', primarily to avoid confusion with the human decision-maker, but also to try to remain agnostic on the question of whether autonomous systems have intentions or intentional states, until such a stage as it would be undeniable. In Chapter 5, when I consider how responsibility *might* come to be shared between humans and autonomous systems, intentionality of this kind is a necessary condition. I simply do not assume this richer property to be true of them until such point as it would be irresistible to ascribe it.

For the most part, I also avoid using the term 'agent' in reference to an autonomous system (again, until Chapter 5). To be clear, though, 'agent' is also a well-established term of art in

[2] This distinction between positive and negative autonomy is particularly worth making in the context of an inquiry into the locus of moral responsibility for harms caused by autonomous systems. As I discuss in Chapter 3 (§3.3), the common thought in the responsibility gap literature is that such a gap only arises when the system sits in "an uneasy conceptual space" between positive and negative autonomy, or when it is implicitly understood to be positively autonomous (Nyholm, 2018; Sparrow, 2007; Matthias, 2004). By contrast, I posit that the challenge to the Standard View could be equally well expressed when the machine is understood only to have a high degree of negative autonomy.

AI, for goal-directed systems that are self-contained, embedded within environments, interactive through sensors and actuators, and can execute a specific task on behalf of a user (Russell & Norvig, 2003, p. 27; Wooldridge, 2020, p. 138). Moreover, while philosophers of action have a focal interest in intentional human action and tend to reserve the term 'agent' for intentional agency, 'agent' also denotes the very general category of something that does something or acts, and which is causally responsible for change or for the prevention of change (Hacker, 2007, pp. 123 - 125). In other words, an 'agent' is anything that has causal power, and this includes certain inanimate entities; examples of such are "acid or radium, a volcano or the sun" (Alvarez & Hyman, 1998, p. 244). Agents with the capacity to act intentionally are, on this more liberal account, a sub-set of the far more general category of entities that have causal power. Thus, though I speak of 'autonomous systems' and reserve 'agents' for human agents, I do think it is true that today's systems are minimal, causal, goal-directed agents, in line with this liberal conception.

1.1.1 Machine autonomy

To lay some conceptual groundwork, I now elucidate the concept of machine autonomy. My purpose is to create a structure for thinking about the capabilities of autonomous systems, providing a reasonably fine-grained account of their relation to human control. Machine autonomy is defined (negatively) in terms of independence from the human user or operator: "a system is said to be autonomous with respect to human control to a certain degree" (Müller, 2020, p. 10). During its operation, there is no or limited need for direct, human control. Indeed, this falls out of its agentive function. As Di Nucci argues, the idea is that this very loss of direct control will be compensated for by the enhanced performance or other strategic or economic benefits it confers (Di Nucci, 2020, p. 130). In these sub-sections, I analyse the concept of machine autonomy in terms of three respects in which a machine can be independent of human control to a certain degree. I call these 'the three dimensions of machine autonomy'. As shorthand, I abbreviate them to the x-axis, the y-axis, and the z-axis, respectively.

The first of the three dimensions is the machine's operational independence from direct human interference or intervention. This is what I call the 'x-axis' of machine autonomy. This is the dimension of machine autonomy that people generally have in mind when speaking of autonomous systems. It is captured in the influential SAE Levels of Autonomy

for driving automation, for example, ranging from driver support at SAE Level 0 to full replacement of the human driver at all times at SAE Level 5 (SAE, 2021).

Fully autonomous systems on the x-axis carry out all elements of a whole task independently of human interference, such as the Level 4 or 5 self-driving car described above. *Partially autonomous* systems on this axis carry out parts or elements of tasks independently of human interference. Systems with partial autonomy are of two kinds. First, those whose outputs are only a discrete *part* of the full task, such as an SAE Level 3 specified vehicle with an autonomous lane-keeping capability, which only has full control when the vehicle is on the motorway. Second, those which complete some *elements* of whole tasks without human intervention (e.g. classification or prediction), but a human user makes the final decision whether to act on those outputs and implement them in the world. The paradigmatic examples here are algorithmic decision-support systems, such as those used in credit-scoring, loan or prison sentencing, and bail decisions. Despite their manifold ethical implications, this thesis is not concerned with this latter kind of support system, since my focus is moral responsibility for harms caused *directly* by autonomous systems. The x-axis of machine autonomy – the system's operational independence from direct human interference – ranges from the system autonomously undertaking *some parts* or *elements* of a task to its doing so for *all parts* and *elements* of a task.

The second dimension of machine autonomy is independence from continuous human oversight. This is what I call the 'y-axis' of machine autonomy. While the agentive function of autonomous systems is to replace the human in the execution of a complex task, there is usually an expectation that its deployment will at least be monitored by a human *in situ* or remotely by a human somewhere in an operations or command centre. The degree to which the system is independent from constant or continuous human oversight (some parts of this oversight might well be delegated to another computational system) is what determines its degree of autonomy on the y-axis. The two dimensions represented by the x- and y-axes are conceptually separable. A machine that is low on the x-axis, only carrying out *part* of the whole task (I have abstracted out from this inquiry those systems that perform only some *elements* of task) could do so with no or very little human oversight, and therefore be high on the y-axis. Conversely, a fully autonomous system that is high on the x-axis may be subject to the constant monitoring and scrutiny of a human operator, and therefore be low on the y-axis.

While conceptually separable, however, human operational intervention is practically dependent upon oversight. One cannot intervene effectively if one has not been paying attention. And systems with autonomy in the mid-range of the x-axis, such as SAE Levels 3 and 4, are deployed with commensurate duties of oversight on humans-in-the-loop – who will need to step in when the system has completed its part of the task, as well as be alert to a transition demand in the event of unexpected events.[3] We can think of the x-axis as the degree to which there is a 'human in-the-loop', by which I mean a human intervenes or stands by to intervene, and the y-axis as the degree to which there is a 'human on-the-loop', by which I mean a human monitors and oversees (European Commission, 2019, p. 16; Merat *et al.*, 2019, p. 92).

As an empirical point, it should be noted that it is psychologically and physically taxing for a human to remain on-the-loop for extended periods of time (Merat *et al.*, 2019). It is difficult to maintain the concentration required for continuous oversight. There is also a well-documented human tendency towards automation bias, such that humans-on-the-loop may over-estimate the accuracy of the system and fail to scrutinise these closely.[4] Moreover, even a conscientious human overseer may, particularly with certain ML-based autonomous systems (as we shall see in the next sub-section, and in more detail in §3.3.2), have limited epistemic access to the system in full throttle. There are limits to the depth of human oversight.

1.1.2 Decision-making function

The third dimension of machine autonomy is independence from explicit human instruction. This is what I call the 'z-axis' of machine autonomy. By 'explicit human instruction' I mean that human engineers clearly and directly encode the precise steps by which the system will

[3] This emphasis is clear in evidence in the UK and Scottish Law Commissions' development of regulations for the safe deployment of highly automated vehicles (Law Commission, 2020; Law Commission, 2019).

[4] Automation bias was in evidence in several plane crashes in recent years, such as the Asiana Airlines flight 214 which crashed in 2013 and killed three people, and in which overreliance on automated flight features played an important role in the crash, and the Air France flight 446 crash in 2009 which was partly due to the crew's confusion and reliance on faulty airspeed measurements (Borenstein, Howard & Wagner, 2017, p. 127). By the same token, Stanislav Petrov, who famously averted a nuclear war by refusing to accept a system's output as accurate, represents the antithesis of a person prey to automation bias. Anecdotally, I have been told that air traffic controllers display similar scepticism and scrutiny of a system's results.

arrive at the desired output. This contrasts with more implicit methods employed by engineers by which the system is set up to learn or infer how to arrive at the desired output.

At the start of this Chapter, we saw that a defining feature of autonomous systems is their fulfilment of a decision-making function. We can now be more specific. As a matter of fact, the decision-making function of today's autonomous systems is increasingly achieved by the implementation of computational techniques that involve a move away from explicit human instruction towards more implicit 'guiding' of the system towards the desired result. I identify this move with the trend away from classic, rule-based techniques towards learning-based techniques. In this sub-section, I describe those computational techniques briefly, both to furnish a deeper understanding of the systems that can directly cause harm, and the internal mechanisms by which they would do so, and to illustrate the sorts of systems that would manifest increasing autonomy on the z-axis.

For most of the history of AI, the rule-based paradigm has been dominant. The central insight of this paradigm is that computational systems will best do what it takes human intelligence to do if they are directly programmed with logic-based algorithms that explicitly specify the steps between input and output. The typical system in this paradigm – the expert system – is encoded with logical rules to manipulate symbolic representations of expert knowledge in the relevant field, such as medical diagnosis.[5] An expert system's operation is governed precisely by the rules yielding output from input. Such systems, and those in this paradigm generally, are sometimes called 'deterministic systems': their outputs are uniquely determined by their inputs and initial state. Amongst other things, this makes it possible to verify, deductively, that the system conforms to a well-defined specification and hence will be safe in its intended context. It also makes the system's outputs easily interpretable. But despite their virtues, rule-based systems are only suitable for tasks and environments that *are* specifiable in this way. Many tasks that humans do quite naturally, such as recognise faces in a crowd, or *learn* to do quite naturally, such as drive through a populated area, are not so amenable to explicit rules and instructions. In many cases, it is not clear what the rules for such functions would even be.

Another way to perform successfully on these tasks is to build systems that can infer or extrapolate the rules from data. This leads us to today's dominant paradigm: Machine

[5] See, for example, Buccanan & Shortliffe (1984)

Learning. Rather than encode rules for the whole input-to-output process, algorithms in this paradigm are general purpose, initial learning algorithms, or 'meta-algorithms' that are applied to large datasets (Kearns & Roth, 2019, p. 6). The resultant inferences from the data become the rules the system applies to new input data to determine an output. The fully trained result is the model. The model space – the internal structure of the model – is generally (but not always) an artificial neural network (ANN). ANNs are abstract structures that mimic, very loosely, the neurons (nodes) and the excitatory and inhibitory synapses (connections) that play a functional role in the brain. They have three kinds of layer: the input layer; the hidden layers; and the output layer. The input layer receives data from outside the network. Connections from the input layer to the first hidden layer assign weights to each input datum. Nodes in the hidden layers are sites of simple, single computations that take as an input the value of each of the previous nodes with their connection weights. The subsequent outputs then serve as inputs for the next hidden layer of nodes, until the process reaches the output layer, which transmits the final outputs to the system's actuators, such as the self-driving car's wheels, which effect the output concretely in the world.

The learning algorithms chosen will depend on the learning problem category, such as classification, clustering, prediction, planning, or optimization (improving on a machine's performance of a task), as well as the environment in which the system is to be deployed (Russell & Norvig, 2003, pp. 32-45). The most common of these is *supervised learning* algorithms. Supervised learning algorithms are generally used for classification tasks and for prediction tasks. Here, labelled training data, such as images, are fed to the machine together with the desired output, on the basis of which it computes a vector of numerical scores. An 'objective function' – the system's goal – is written which aligns the 'correct label' with the highest score. The learning algorithm normally employs a form of gradient descent: modifying the weight of the internal parameters to reduce the distance between the system's output scores and the highest score (LeCun, Bengio & Hinton, 2015, p. 437).

Another form of machine learning is *unsupervised learning*. Here, the system extrapolates rules and patterns from unlabelled training data. Typically, the system detects characteristics of the data that make them more or less similar to one another. Unsupervised learning algorithms involve distance metrics or density estimation, and are generally used for clustering and data-mining tasks. The implementation of unsupervised learning algorithms would push a system

further up the z-axis than supervised learning algorithms, because there is even less explicit instruction of the system.

Both supervised and unsupervised learning algorithms are good for finding pattens or hidden structure in data. The third main form of machine learning, *reinforcement learning*, follows a different approach. With reinforcement learning, a system is 'let loose' in an environment. It is not explicitly given the desired output, but it is programmed with a goal, namely its 'reward function' – again, the system's goal – which it is designed to optimise through an iterative trial and error learning process. The reward is generally delayed: the output does not just affect the immediate reward, but also the next situation, and hence future rewards. In terms of the explicitness of instruction, reinforcement learning sits somewhere between supervised and unsupervised learning. Reinforcement learning algorithms are used for tasks that require navigation, motion control, planning, and interaction, but they have also been highly successful in gaming, most notably Deep Mind's AlphaGo which in 2016 beat the world champion at the ancient, and notoriously difficult, Chinese board game, Go.

Though such techniques are recently more dominant and have driven a great deal of the current success in autonomous technology, Machine Learning and ANNs are by no means a younger research area. Since the very earliest days of AI, inspiration has been taken from psychology, cognitive science, and biology (Wiener, 1948). But several breakthroughs in the late 1980s and the 1990s increased their potential for success. First, the back propagation algorithm, whereby the weights between nodes in the network can be iteratively altered to minimise approximation errors in the output (Rumelhart, Hinton & Williams, 1986). Second, increases in processing speed, achieved through improvements to the semiconductors and processing units on which the networks run (Moore, 1965).[6] Third, a massive increase in the quantity of available data, due to the Internet and increased data storage capacity. But it was the combination of these breakthroughs with the development of deep neural networks (DNNs), which began around the year 2006, which brought on the new wave neural network research (Wooldridge, 2020, pp. 184-185).

[6] Since 1965, the rate of increase in processing power has often been called 'Moore's Law', following a paper by Intel co-founder Gordon Moore which posited that this power roughly doubles every two years. It should be noted, however, that this is not a law of physics, but a hypothesis about the rate of technological progress, and there is no reason to expect that it will continue on this trajectory indefinitely.

Deep neural networks combine three features; in short, they have "more layers, and more, better-connected neurons" (Wooldridge, 2020, p.185). Deep learning uses networks that have up to hundreds of hidden layers, each of which can process data at a different level of abstraction, up to a million nodes, enabling much more powerful computation at each level of abstraction, and up to hundreds of millions of connections between nodes, each with adjustable weights. After the training and learning phase, developers validate and test the model, to evaluate its performance on unseen or novel data against pre-defined criteria, such as the number of false positives amongst the classifications, and to refine this performance until it is satisfactory, and the developer is confident the resulting model will generalise effectively over intended future use cases.

But in overcoming the disadvantages of explicit, rule-based programming, today's most powerful Machine Learning models also lose the advantages of that paradigm. As Sullivan notes, DNNs are not *completely* inscrutable 'black boxes'. Software engineers choose the meta-algorithms, the activation functions, the data, the number of layers and nodes, and they also test the subsequent model. But DNNs *are* inscrutable at the crucial implementation level (Sullivan, 2021, pp. 17-18). That is, the different weights they learn and assign for each data point cannot be directly inspected or known. Moreover, given the sheer complexity of its processing, it is not technically possible to interpret precisely how the model determines the output, although 'Explainable AI' techniques have been developed to provide higher-level approximations of the system's logical and causal process, and the features it has identified as salient in the model.[7] Even so, these features may not match features humans would use in the course of coming to a decision, "so interpretation would remain difficult even if the model could be inspected" (McDermid *et al.*, 2021). Many systems that have increased autonomy on the z-axis are also extremely complex. This, in turn, makes them more difficult for human agents, even experts, fully to understand or robustly to predict their outputs.

For the sake of completeness, there are two further kinds of learning technique that should be mentioned. The first is Bayesian networks, which model probabilistic relationships. Bayesian techniques have long been popular for encoding uncertain human knowledge in Expert Systems, but today they are also being used to derive probabilistic statistical inferences from data. These are particularly useful for prediction tasks, and many Bayesian networks also represent causal relationships between variables (Pearl, 1988). Unlike DNNs,

[7] For a good discussion of Explainable AI (XAI) techniques see Arrieta *et al.* (2020)

though they are more implicitly programmed than rule-based systems, Bayesian networks *are* still readily interpretable by the human expert.

The second further kind of learning system is constructed abstractly to represent the biological processes of experimentation, adaptation, mutation, and fitness selection. Here, Artificial Intelligence meets what is sometimes called Artificial Life (Boden, 1996). One important computational technique in this paradigm is the implementation of evolutionary and genetic algorithms, which are typically used for search and optimisation problems, and to make heuristic, or 'rule of thumb,' inferences mapped to a fitness function (Holland, 1992). The fitness function is similar to the objective function of supervised learning algorithms or to the reward function of reinforcement learning algorithms, insofar as it constitutes the system's goal or target. Evolutionary algorithms are good for tasks "where the correct, or even good, solutions and not known, but need to be discovered" (Mikkulainen, 2021, p. 1). This includes more creative tasks, such as engineering design, as well as complex logistical tasks, such as multi-dimensional scheduling. Here, too, system performance is achieved through techniques that more implicitly guide the system to a result rather than directly specify all the steps it should take to get there. At the same time, evolutionary algorithms are also 'black boxes' at the implementation level; the pathways by which they achieve results are largely inscrutable *post-hoc*.

Finally, learning systems can be either what is called 'offline' or 'online' (Royal Society, 2017, p. 20). Offline learning systems are fully trained and validated and then 'frozen' before deployment. Any changes to the model are made by a human. The majority of today's autonomous systems incorporate offline supervised machine learning models. As ML-based systems, these have more autonomy on the z-axis than rule-based systems. However, it should be noted that, in being fixed before operation, and not *continuing* to learn, the autonomy of such systems on the z-axis sits is lower than with online learning systems. Online learning systems are trained before deployment, but continue to learn 'in the wild,' updating their models in real-time in reaction to feedback from environmental stimuli. Online learning systems are adaptive. Online learning systems would be far more highly autonomous on the z-axis.

I have defined the z-axis of machine autonomy as that dimension of independence from direct human control which tracks the systems' increasing independence from explicit human

instruction during the engineering phase. I identify 'explicit human instruction' with a system's being rule-based in the classic sense. This contrasts with more implicit, learning-based techniques which guide a system to a desired outcome. One of the advantages of the rule-based paradigm is the straightforwardness of verification and validation. As we have seen, many learning-based systems are difficult to interpret at the implementation level. Many are also inherently uncertain. How does one gain confidence in the outputs of such a system (verification)? And how does one demonstrate that they reliably align with the intentions of engineers and users (validation)? As Burton *et al.* put it:

> *"Machine learning functions do not deliver clear-cut answers. For example, for a given video frame, they might classify the probability of a pedestrian inhabiting a certain portion of the picture as 83%, but in the very next frame – which for humans is imperceptibly different to the last – they may "misclassify" the same object as only 26% probability of being a human and 67% probability of being a road sign. In addition, the processes which lead to these decisions are difficult to decipher. These attributes result in a paradox or "no free lunch" effect, where the problem of deriving a suitable specification of the intended behaviour is instead transferred to the problem of demonstrating that the implemented (learned) behaviour meets the intent."*
>
> (Burton *et al.,* 2020, p. 4)

Whereas the x-axis and the y-axis primarily refer to independence from the human *operator*, the z-axis, in referring to increasing detachment from explicit human instruction in the engineering phase, primarily refers to independence from the human *engineers or developers* of the system.

Thus, my analysis of the three dimensions of machine autonomy, reveals that, when we speak of a system's independence from direct human control, this is both multi-faceted and inextricably concerns both human agents on the ground and human agents in the development phase. This underscores my later argument for shared human responsibility.

1.1.3 The three dimensions of machine autonomy

Since, in the Chapters that follow, I refer back to these dimensions of machine autonomy at various points, it will be useful to enumerate them again here. For ease of mental

visualisation, I have called the three dimensions the x-axis, the y-axis, and the z-axis, respectively.[8]

> The *first* is the system's operational independence from direct human interference, the x-axis. This corresponds to the degree to which there is a human-in-the-loop.
>
> The *second* is the system's operational independence from continuous human oversight, the y-axis. This corresponds to the degree, or depth, to which there is a human on-the-loop.
>
> The *third* is the independence of the system from explicit human instruction, the z-axis. This corresponds to the degree to which the computational techniques that have been used to build the system and move away from the rule-based towards the learning-based paradigm.

Let us consider two of the examples in the Introduction to see how they might be plotted on the three dimensions. The self-driving car, in undertaking the full dynamic driving task independently of Jacinta's direct operational control, is highly autonomous on the x-axis. While Jacinta's monitoring of the car's operation is minimal, she is present in the vehicle and aware of the environment. We can therefore say that Myrtle is moderately autonomous on the y-axis. Many of the vehicle's components, such as its computer vision model, rely on DNNs. It is therefore also relatively high on the z-axis. It would be higher if it continued to learn 'in the wild' but it does not. The assistive robot in the care home operates independently of manual control. When Tom is ready to use it, after starting it up, no further intervention is required by a care worker. In this respect, then, it is also high on the x-axis.[9] Let us imagine that there is a human carer in the room at the time, observing and overseeing the manoeuvres. The system is therefore low on the y-axis. However, let us further imagine that

[8] To note, my characterisation of machine autonomy on these dimensions does not include the system's operating environment, which the SAE Levels of Autonomy do include. I thank Ibrahim Habli for bringing this to my attention. Insofar as environment influences the degree of intervention or oversight, then this could be incorporated by the x-axis and y-axis, respectively. Otherwise, if the nature of the environment does not bear on degrees of independence from direct human control, it is not directly relevant to this point. I should add that my elucidation of machine autonomy is for the purposes of providing the conceptual groundwork for understanding the relation between autonomous systems and human control. It is not intended as an alternative to the established scales and taxonomies, such as the SAE Levels.

[9] However, there is complication here insofar as the assistive robot responds to Tom's cues, so that there is a sense in which its operation is dependent upon human interference, which perhaps lowers, or at least complicates, its autonomy on the x-axis.

the system has been built using includes deep reinforcement learning techniques, and that even after an accident investigation it takes detailed diagnostics to posit the combination of the anomalous user cue and the noise from lawnmower outside as the cause of the system's unwanted and unforeseen behaviour. This makes it highly autonomous on the z-axis.

As I have said, my purpose in elucidating the three dimensions of machine autonomy is to show how the independence of these machines from direct human control is multi-faceted. Greater autonomy on the x-axis and y-axis, both of which refer primarily to increasing independence of the system from the human operator, translates as independence from direct human *operational* control over its outputs and their consequences. Greater autonomy on the z-axis, which refers to increasing independence from the explicit instructions of the human engineer, translates as independence from direct human *rational* control over those outputs and their consequences. Learning-based systems infer how to reach a desired output from the combination of learning algorithm, goal, and data. The intended functionality is therefore implicitly rather than explicitly specified by engineers. This is in many ways an advantage of learning-based systems, since it provides a technical solution for the kinds of functionality that cannot be easily specified using classic, rule-based techniques. But it also creates the possibility that the system will behave in surprising ways. In addition, many powerful learning-based systems are also vastly more complex and less readily interpretable than systems built using rule-based techniques, making their behaviour more difficult to predict or explain which, in turn, affects rational control. As I shall argue in Chapter 3, these features place pressure on the fulfilment of the traditional necessary conditions in our frameworks for ascribing moral responsibility.

It should be noted that my concern is not the existential risk that a future Artificial Superintelligence might pose. At the heart of these worries about existential risk is the notion that once AI crosses a certain threshold of intelligence, humans will not be able to control it all (Russell, 2019; Bostrom, 2014; Chalmers, 2009). Though this thesis also addresses the question of control, it does not engage in analysis of the prospect of an Artificial Superintelligence, nor of its normative consequences.

1.2 Moral responsibility for harms caused by autonomous systems: the Standard View

Given the nature of the tasks that are transferred to them, and the domains in which they are deployed, there is a risk that autonomous systems will directly cause morally significant harm to human agents. In many cases, the risk is of physical harm, but it may also be psychological, economic, social, or political harm. Such harms would not be natural catastrophic events like earthquakes. Autonomous systems have not fallen from the sky; human agency has been, and always will be, involved in the creation and deployment of the systems. It would therefore be reasonable and natural for the victims of the harm or their families, or for society at large, to ask: Who did this? Who deserves blame?

The position that I call the Standard View is the common sense and seemingly obvious position that moral responsibility for harms caused by autonomous systems rests completely and exclusively with human agents. Human moral responsibility has long been "one of the standard operating assumptions of computer ethics" (Gunkel, 2012, p. 27). It is derived from the orthodox, instrumentalist theory of technology that all technological artefacts are mere tools, of human creation and use, which are simply extensions of human capacities, dependent fully upon human intentionality and human control (Gunkel, 2017; Gunkel, 2012).[10] Instrumentalist defences of the Standard View also tend to defend the *exclusivity* of human moral responsibility by emphasizing the capacities that human beings have, which are necessary to being a responsible moral agent, and which machines lack, such as consciousness (Himma, 2009), or the capacity to act for reasons (Purves, Jenkins & Strawser, 2015). As Coeckelbergh expresses the view, "even if AI technologies gain more agency, humans remain responsible since only the latter can be responsible: artificial intelligence technologies can have agency but do not meet the traditional criteria for moral agency and moral responsibility" (2020a, p. 2053).

In the face of this increasing machine autonomy on the three dimensions, the challenge for the Standard View now is not to emphasize that the machines are *not* morally responsible, but to defend the seemingly common-sense conclusion that human beings *are*. This is because the normal necessary conditions for warranted ascriptions of moral responsibility – control and knowledge – are affected as the machine moves further up the three scales of

[10] For expressions of the instrumentalist view, see: Coeckelbergh (2020a, p. 2053); Boden *et al.* (2017); Johnson and Miller (2008); Johnson (2006b); Searle (1997, p. 190); and Dennett (1989, p 298).

machine autonomy. Matthias is to be credited as the first scholar to articulate this worry in the literature (Matthias, 2004). It has been a particular point of debate in respect of autonomous and highly automated weapons (Himmelreich, 2019; Di Nucci & Santoni de Sio, 2016; Schulzke, 2013; Sparrow, 2007) and to a lesser extent in respect of self-driving cars (Hevelke & Nida-Rümelin, 2015). The focus now is on showing that "responsibility will not evaporate" (Santoni di Sio & van den Hoven, 2018, p. 2).

A framework that acknowledges the increasing autonomy of systems and provides principled grounds for upholding the Standard View is of pressing practical concern in the governance of autonomous systems. In the plethora of sets of ethical principles that have been published in the last five years – such as the OECD Principles on AI (2019), the European Commission's Guidelines on Trustworthy AI (2019) or the principles of the Montreal Declaration (University of Montreal, 2018) – the importance of human responsibility has almost unanimously been emphasized and expressed (Fjeld *et al.*, 2020; Hagendorff, 2020; Jobin, Ienca & Vayena, 2019). Despite a tendency to ambiguity in some these documents – particularly between the descriptive and the normative – which is, perhaps, to be expected in this specific multi-disciplinary context (Tasioulas, 2019, p. 58), a philosophical defence of the Standard View position would be consistent with the wider public narrative around responsibility for harms caused by autonomous systems. Of course, this is not yet to say that the Standard View is correct. Nonetheless, a well-formed explanation of its correctness, and a conceptual framework for attributing and ascribing moral responsibility to human agents for harm caused by autonomous systems, would be valuable contribution to the real-world regulation of their development and deployment.

If a defence of the Standard View is going to rest on instrumentalist assumptions, then it needs to show how autonomous systems that are highly autonomous on the z-axis fall within this category of 'mere tools'. Alternatively, a Standard View position may locate moral responsibility with humans – completely and exclusively – without recourse to the instrumentalist understanding. I take this latter route: I argue that they are no longer just our tools, they are our delegates or proxies. Just as we necessarily relinquish considerable rational control over, and epistemic access to, the decision-making of human proxies to whom we delegate tasks, so too do we relinquish this when we delegate tasks to autonomous systems. Moreover, just as principals are responsible for the consequences of what their human proxies do, so are they responsible for the consequences of what their artificial proxies do

(Di Nucci, 2020, p. 184). This situates my account within that growing strand of the Standard View that grounds human responsibility in a delegation relation between human agents and these system (Di Nucci, 2020; Millar & Kerr, 2016; Millar, 2015; Di Nucci & Santoni de Sio, 2014).[11] More specifically, my argument is that the instantiation of a delegation relation between human agents and autonomous systems is logically sufficient for the warranted attribution of moral responsibility to human agents.

The Standard View is capacious. I include within the position theories of distributed moral responsibility – so long as the responsible 'nodes' in the responsibility network or collaborative agential unit are exclusively human (Nyholm, 2018).[12] It is also flexible enough to accommodate the view that human agency radically transforms through interaction with autonomous systems. On my understanding, what qualifies a position as an expression of the Standard View is that humans are exclusively responsible for the harms caused by autonomous systems, and that human moral responsibility coverage is complete: no non-humans bear moral responsibility for harms caused by autonomous systems and there are either no 'responsibility gaps' with respect to those harms, or these gaps can be bridged and are not fatal.

The Standard View can be analysed in terms of two distinctions. The first distinction is between a hard and a soft version of the position. A hard Standard view would maintain that, *necessarily*, human agents are exclusively morally responsible for harms. This would most likely turn on the original defences of the Standard View, such that humans alone have the agential capacities to bear responsibility. A soft Standard view, by contrast, holds that it is only *contingently* the case that humans are exclusively morally responsible for the harms that autonomous systems cause. That is, that it is possible that the autonomous systems themselves *might* be morally responsible for them, in some respect – but that at present, as a matter of fact, they are not. The second distinction is between a monistic and a pluralistic version of the position. Anyone who holds that moral responsibility rests with one class of human agent (for example, with the developers alone), we may call a Standard View monist.

[11] Thoma can also be regarded as implicitly within the delegation framework, given the presupposition that autonomous systems *are* proxies: "we can understand [artificial agents] as acting as proxies for human agents, as making decisions on their behalf" (2021, p. 2)

[12] Loh & Loh (2017) and Floridi (2016) do not, for example, count as a Standard View position. Because they hold that autonomous systems share moral responsibility within a responsibility network encompassing a wide range of other agents, they move beyond the exclusivity and comprehensiveness claims I have imputed to the Standard View.

Where it is maintained that moral responsibility rests with more than one class of human agent, we may call this position-holder a Standard view pluralist.

The position I defend in this thesis is a soft, pluralist Standard View position.

Summary of Chapter:

I have elucidated the concept of machine autonomy as having three dimensions: independence of the system from direct human interference; independence of the system from continuous human oversight; and independence of the system's behaviour from explicit human instruction. For short, I call these the x-axis, the y-axis, and the z-axis of machine autonomy, respectively. The x- and y-axes primarily concern independence from the low-level users on the ground and the z-axis primarily concerns independence from the high-level designers and engineers upstream in the development process. Increasing autonomy on these dimensions puts pressure on our traditional conceptual frameworks for attributing moral responsibility to humans for harms caused by autonomous systems, which common sense tells us is where it belongs.

CHAPTER 2
THE *LOCUS OF MORAL RESPONSIBILITY* QUESTION

In the second Chapter, I formulate the central question that a defence of a Standard View position needs to answer. This is what I call the 'locus of moral responsibility' question, which asks the grounds for ascribing moral responsibility to particular human agents for harms caused by autonomous systems. This can be expressed as a three-term relation: S is morally responsible for X in virtue of P. First, I restrict the scope of X to unforeseen harms, and show the pathways by which such harms might come about. Second, I clarify my conception of moral responsibility and set out two stages to my procedure for answering the question: to determine the locus of moral responsibility as attributability, first, and the locus of moral responsibility as blameworthiness, second. The locus of moral responsibility question will be answered once both parts have been addressed.

2.1 The unforeseen harm

The agentive function of an autonomous system is to replace a human in a task – such as driving through city streets, providing physical assistance to vulnerable people, or destroying military targets – that has historically required human decision-making, interpretation, and judgement. The system's substitution of a human decision-maker, which widens the scope of possible hazards and increases the moral significance of the situations it is likely to face, combines with the synchronic and diachronic complexity of the environments in which it is deployed, to create conditions of uncertainty about the impact of the system's real-world behaviour. The *locus of moral responsibility* question concerns responsibility for harms that occur at the limits of foresight in these conditions of uncertainty.

Decision theorists often make a distinction between decision-making under risk and decision-making under uncertainty (Hansson, 2003, p. 293). Decisions made under risk assign strict probabilities to a set of possible outcomes. But in decision-making under uncertainty, probabilities are not known with precision or at all. It is a common strategy in decision theory to reduce uncertainty to probability, by assigning exact probabilities to all uncertain events (Harsanyi, 1977, p. 381, cited in Hansson, 2003, p. 307). Nonetheless, the conditions under which, and for which, decisions about autonomous systems and risk are

made are conditions of uncertainty. As such, there may be unforeseen harms that are not captured by the accepted risk thresholds – the amount of risk that is acceptable or tolerable – as I explain in more detail in my discussion of the various possible pathways to unforeseen harm below. These are the unforeseen harms with which this thesis is concerned.

I make three further clarifications. First, by 'unforeseen harm' I mean actually unforeseen rather than unforeseeable. Moreover, I have in mind that its being unforeseen is not necessarily negligent. We might call this *reasonably* unforeseen. In practice, how this would be determined is vague. People have different dispositional and occurrent capacities for foresight. And, in the context of new technological systems, what it is reasonable to expect people to foresee, and to foresee rightly – whether in the factual or moral sense of 'right' – is by no means clear. Moreover, it may not be possible to quantify or predict with reliability *ex ante* the likelihood or severity of certain harms or risks of harm. Even in cases where risk *can* be calculated as a probability in advance, for example on the basis that a system's predictions have a known false positive rate, it is not clear – yet – what constitutes the morally or societally acceptable threshold of risk (Law Commission, 2020). Noting these difficulties and conditions of uncertainty, however, the harms I delineate as being of principal concern are those that the relevant humans sincerely did not foresee and would have mitigated or taken pains to avoid had they correctly anticipated them.

Second, the harm is a consequence. This consequence is logically but not causally separable from the action (or rather, output); it is extrinsic to the action (Von Wright, 1963, p. 116). Our question therefore concerns responsibility for consequences rather than just for system outputs. This increased distance of the harm from human agents further occludes ascriptions of moral responsibility. After all, as von Wright states, if a person "does not foresee the consequence (or at least realizes the 'serious possibility' that it will happen), then he can, in a sense, not even be rightly said to have done the consequent thing." (1963, p. 124). To adapt Anscombe's classic example (1957, p. 37), if the villagers were poisoned as a consequence of the pumping of the well, and if the pumping agent had no idea or intention that her pumping would cause this poisoning, even though she is morally responsible for pumping the water, she is not clearly morally responsible for the poisoning of the villagers. Some of the harms for which we seek to locate moral responsibility are akin to this poisoning of the villagers – in fact, this is not dissimilar to the hypothetical example I gave in the Introduction, of the missile that contaminates a village's water supply – but our case is complicated by the fact

that the causer of the harmful consequence is an autonomous system and not a human, by most accounts falling short of the agential capacity required to warrant an ascription of moral responsibility.

Third, I restrict the focus to an unforeseen harm caused directly caused by the system and not by the human-in-the-loop (see §1.1.1). This restriction excludes AI and ML-based decision support systems from the inquiry, such as the notorious COMPAS algorithm (Larson, 2016), or systems used by doctors in medical diagnosis (Habli, Lawton & Porter, 2019). My reason for omitting these cases is that these are probably directly attributable to human agents. As a matter of fact, the reliance upon the human-in-the-loop could place an undue legal and moral burden on her; in practice, she may be unable to live up to her obligation always to intervene effectively or to scrutinise the system's predictions in great detail.[13] In cases where the human intermediary is no more than a rubber stamp for the system's outputs, much of my analysis here will apply. But formally I restrict my inquiry to harms that are directly caused by the autonomous system. The point is that immediate causal antecedent of the harm is the autonomous system.

The output and the consequence are logically (but not causally) separable. In respect of the foreseen/unforeseen distinction, there are four possible causal pathways that could lead to the harm, which are set out in Table 1 below. Rows II and IV are the unforeseen harms with which we are primarily concerned. These are abbreviated as UHC. Foreseen harmful consequences are abbreviated as FHC. I present the table below and provide some explanation and clarification afterwards.

[13] For an excellent discussion on the need for human agents to avoid doxastic negligence and to exercise deliberative capacities here, see Zimmermann & Stronach (2021)

Table 1

	Human high-level decision to develop or human low-level decision to deploy the autonomous system	The autonomous system's output	The harm caused by the autonomous system's output
I	Voluntary	Foreseen	Foreseen (FHC)
II	Voluntary	Foreseen	Unforeseen (UHC)
III	Voluntary	Unforeseen	Foreseen (FHC)
IV	Voluntary	Unforeseen	Unforeseen (UHC)

This first column of Table 1 describes the natural human candidates for moral responsibility. If moral responsibility for unforeseen harms (Rows II and IV) is to rest with any human agents, the developers and deployers are, at least, the obvious candidates to start with. For the rest of the thesis, adapting a distinction drawn by Thoma, I call these the *high-level* and the *low-level* human agents, respectively. Thoma draws the distinction thus: low-level agents are "the individual users of the artificial agents, or the kinds of individual human agents artificial agents are usually replacing"; high-level agents are "designers, distributors or regulators, that is, those who can potentially control the choice behaviour of many artificial agents at once" (2021, p. 2). Without the high-level agents, the systems *could* not be deployed; without the low-level agents the systems *would* not be deployed. While high-level agents render possible the use of the systems, low-level agents actualise it.

In the first column, I also stipulate that the high-level and low-level human decisions are voluntary. This means that none of the human decision-makers involved are subject to force, duress, or coercion in their acts in this column. This may, in fact, be optimistic. Individuals in powerful corporations may have a small voice and a limited ability to refuse to do what is asked of them. Moreover, as these technologies become more embedded in our social, political, and economic structures, individuals may strictly be able to 'opt out' of deploying or using autonomous systems only at severe cost or inconvenience to themselves. Nonetheless, for the purpose of this inquiry, let us assume that high-level and low-level human agents do make voluntary decisions.

A Standard View position-holder would be inclined to say, I think, that moral responsibility for all harms caused by an autonomous system simply traces back to the high-level and low-

level agents making voluntary decisions in the first column, at the very start of the causal pathway. But what is *not* obvious – and what a Standard View account needs to answer – is on *what grounds* such an ascription would be warranted. The fact that antecedent decisions were voluntary on the part of human agents does not mean that the consequences were voluntarily chosen, it is not clear that voluntary antecedent decisions alone are sufficient for an ascription of moral responsibility. As Floridi argues, intention is not closed under causal implication. It is not the case that, if S means to cause *a*, and *a* causes *b*, S means to cause *b* (2016, p. 4). The same applies to voluntariness. It is not the case that, if S chooses to cause *a*, and *a* causes *b*, S chooses to cause *b*. A deeper explanation is required to ground moral responsibility with S. My answer to the *locus of moral responsibility* question, turning on the delegation relation, provides this deeper explanation.

The second and third columns of Table 1 capture the causal pathway from the system's behaviour to its consequence in the world. The 'foreseen' or 'unforeseen' is meant to denote foreseen by *any* of the humans, whether high-level or low-level. The 'or' is inclusive. That is, we can rule out cases where either the low-level agent foresaw the harm and wittingly used the system contrary to the engineer's intentions, or where the high-level agent foresaw the harm and wittingly withheld that information from the low-level human agent. In practice this may also be artificially optimistic, but it is my working assumption.

Given this emphasis on unforeseen harms, and specifically those that were non-negligently unforeseen, a thought at this juncture might be that responsibility rests with high-level or low-level human agents in the first column of Table 1 as a matter of strict moral liability, and that this constitutes the most obvious and best defence of the Standard View. Strict moral liability can be understood as the moral counterpart of strict liability in law, where agents are morally obligated to make restitution to those harmed regardless of whether they are culpable for the harm or the behaviour from which it resulted (Capes, 2019).[14] But understanding strict moral liability in terms of making restitution is to get ahead of the question. I clarify later in this Chapter that the conception of moral responsibility I am seeking is that of *being* morally responsible, or responsibility as attributability - being rightly open to answer for a harm in the first place. It is a basic assumption amongst moral philosophers that attributability is necessary for being justly *held* morally responsible, which is called

[14] It should be noted that strict moral liability is itself a contested notion. Hart, for example, denies that a person is morally obligated to redress harms for which the person is not morally to blame (Hart, 1961, p. 168-9).

responsibility as accountability (Watson, 2004, p. 263, 278). But strict moral liability, understood as obligation to make restitution for a harm, is an aspect of responsibility as accountability. A strict moral liability defence of the Standard View goes straight to the practice of *holding* responsible irrespective of whether the accountable or liable agent actually *is* responsible in the first place. Indeed, as Shoemaker notes, the primary reason legal philosophers such as Hart find the notion of strict legal liability morally unjustified (even if justified on pragmatic grounds) is precisely because it "undermines the presumed entailment relationship between accountability and attributability." (Shoemaker, 2013, p. 160).

Alternatively, a strict moral responsibility account one might hold that responsibility rests with high-level or low-level agents in the first column of Table 1 as a matter of strict moral attributability. Strict moral attributability can be understood as the phenomena that obtains when harms are always attributable to certain agents in way that makes them rightly open to answer for the harm and to be subject to moral appraisal for it, regardless of whether they non-culpably failed to foresee the harm (Duff, 2009).[15] The question then arises: in virtue of *what* would strict moral attributability be warranted in cases of unforeseen harms caused directly by autonomous systems?

One option could be that this attributability is warranted because autonomous systems are tools, and high-level and low-level human agents in the first column of Table 1 are toolmakers and tool users. The thought here might be that the nature of entity that causes UHC, or the nature of the tasks or roles of toolmakers or tool users, is sufficient to expect them to answer for the harms, even if those harms were (reasonably) unforeseen and fell outside the accepted risk thresholds. This approach would be commensurate with the instrumental theory of technology. To some extent, it has a legal analogue in the form of strict product liability (noting for one thing, that what has been described is attributability not liability, and for another, that product liability in law applies only to manufacturers or retailers, and so would not include tool users within its scope).[16]

An alternative option is to advance the view that autonomous systems are not merely tools, and that their agentive function and their increasing independence from direct human

[15] This is what Duff (2009, pp. 304-308) calls 'strict moral answerability,' but I have used the term 'attributability' for internal consistency.
[16] Indeed, the fact that product liability does not cover harms incurred by tool users is one of the reasons a liability gap emerges with autonomous systems (see Burton *et al.,*2019, p. 11).

control makes them proxies or delegates. On this account, the nature of the delegation relation between some of the high-level and low-level human agents in the first column of Table 1 and autonomous systems is sufficient to expect them to answer for the harms, even if those harms were (reasonably) unforeseen and fell outside the accepted risk thresholds. Moreover – and this is the argument I advance in my delegation account in Chapter 4 – the relation is normatively asymmetrical in such a way that weights responsibility with the principal. Therefore, on the delegation account, the responsibility attribution would be deserved, whereas the toolmaker/tool user account is more arbitrary. To some extent, this delegation framework has a legal analogue in the form of vicarious liability (noting, again, that what has been described is also attributability not liability). Though my concern in this thesis is moral and not legal responsibility, I consider this connection in a little more detail in Chapter 6.

Let us now take a closer look at the different possible pathways to the harm. My purpose in introducing Table 1 is to show, in broad terms, the different ways in which harm, whether foreseen or unforeseen, might come about. While the focus is on the unforeseen harmful consequences of Rows II and IV, it will be helpful to understand these cases within their broader context. I therefore give a survey of Rows I – IV below.

Row I cases

In Row I cases, human agents foresee both the system's output and its harmful consequence. Some military autonomous weapons systems, as well as cyber malware systems, clearly fall into this category, since their purpose is precisely to cause damage. Many of the algorithmic decision-support systems I have strictly excluded from this inquiry, on the grounds of being advisory and not replacement systems, also serve a purpose which has the consequence of constraining a person's liberty or solvency, for example. The infamous COMPAS system predicted a person's risk of recidivism for the purposes of guiding a judge making a bail or parole decision (Larson, 2016). Facial recognition systems are deployed for the purpose of arresting offenders, admitting visitors to prison, allowing people to cross borders. The MIDAS system is used to detect fraud and automatically to demand repayment for it

(Charette, 2018). Being refused bail, or being refused admittance to see a loved one, or being on the receiving end of a bailiff's demands, are harms to the individual affected.[17]

Harms caused by autonomous systems that fall within known and accepted risk thresholds we can count as belonging in the Row I category. If, for example, a system has a false positive rate of x% and a false negative rate of y%, harms caused directly by either a false positive or a false negative output consistent with that outcome probability would be foreseen, even if not desired. Within this category of Row I cases are also outright cases of negligence, "conduct which fails to conform to a required standard" (Turner, 2019, p. 84). The example I gave of the child and Jacinta's self-driving car can be adapted to illustrate. Imagine, for example, developers of the vehicle overlooked that fact that dappled sunlight would lead to accidents, perhaps because meeting an urgent product deadline was their primary focus. This would be a Row I case of negligence.

Row II cases

But our concern is primarily the unforeseen harm. In Row II cases, the output is foreseen but the consequent harm is unforeseen. One example of such a case is the harm caused by what I call an 'honest mistake', when high-level or low-level human agents do not realise that foreseen system behaviour will be harm-causing or they do not realise that they have not specified the system's requirements correctly. Morgan's 'Scholar Bot' can supply a characterisation. The high-level (or low-level) human agents might have foreseen that the autonomous system would provide detailed information in response to questions asked of it – perhaps it is designed to provide a level of detail according to the abilities of the student, and Morgan is a very bright – but not that information at a certain level of granularity and detail would be psychologically damaging in some cases. This, plausibly, would constitute an honest human mistake. Of course, with such cases it will be difficult to determine whether the consequence really was *reasonably* unforeseen. These difficulties may arise with respect to factual knowledge (was it was reasonable *not* to anticipate that the bot would start providing detailed information of horrific torture?), but they might concern moral knowledge too (was it reasonable *not* to anticipate or appreciate the moral gravity of the effects of such information on young adults?). One way to capture these difficulties is to include what I call

[17] Human agents in the first column of Table 1 might say these are harms, insofar as they are injurious to the individual, but they are not wrongs, since they are morally justified, but – for the purposes of our inquiry – it is nonetheless the case that the harm is foreseen.

'benefit of the doubt' cases under Row II cases of honest mistake. As a methodological point, I charitably allow that the harmful consequence that human agents did not foresee was reasonably and rightly not foreseen, and that failing to consider the age of the student, for example, or other appropriateness conditions, did not fall below due standards of care. In other words, the humans are charitably allowed to call their mistakes honest, and not ones that would put them in the Row I cases of negligence. Honest mistakes could well undermine the previous determinations of acceptable risk, because they would be unknown possible risks when the calculations are being made.

Another example of a Row II case is the 'genuine accident', when the harm, while directly caused by the autonomous system, occurs largely because of other unanticipated events or factors in the environment or in the causal pathway from output to consequence. The autonomous missile example was plausibly a genuine accident. It caused the harm, but there was an unanticipated factor because no one foresaw the village's clean water supply was so critically situated within the system's sphere of influence. As before, a claim that a harm was a genuine accident would be subject to scrutiny as to the reasonableness of the lack of foresight. One might point to the quality of the military intelligence supplied to the autonomous missile, and whether intelligence officers should have been more diligent when making their assessments of the area. But some genuine accidents will happen, where the autonomous system is caught up in unexpected events in which its outputs, for some reason, directly (or, in this case, strictly, *indirectly*) injure or wrong a human. And, as above, I allow 'benefit-of-the-doubt' cases with respect to possible negligence in such decision-making. Genuine accidents of this kind may reveal the risk thresholds to be incomplete, if they fall outside of the scope of the risk calculations.

To summarise, in Row II cases of unforeseen harm, we have honest mistakes, where the harm-causing nature of some of the system's behavioural properties was unanticipated, and genuine accidents, in which unanticipated factors in the operating environment, in the causal pathway from output to consequence, were one of the main reasons for the unforeseen harm. And we also have benefit of the doubt cases, where we accept, for the sake of argument, that the harm was the result of a mistake that *was* honest or the upshot of an accident that *was* genuine.

Row III cases

In Row III cases, the output is unforeseen, but the consequence is foreseen. It involves the autonomous system unexpectedly having the desired (or, at least, accepted) harmful effect. The output may be unforeseen (as it would also be in a Row IV, as we shall see below), due to an undetected design or operational error, or due to emergent behaviour. It is difficult to conceptualise a Row III case, unless we understand the 'foreseen consequence' to have a very broad extension to include all possible future harmful consequences of a general type.

It is easier, perhaps, to imagine the analogous positive case of an unforeseen action causing an anticipated *beneficial* consequence. This is what happened with Deep Mind's Alpha Go, whose emergent actions were novel, winning moves in the ancient strategy board game Go against the world champion, Lee Sedol. Some of the moves were so novel they had never been seen by human players of the game before. This was due to its powerful deep reinforcement learning model. This example shows how systems with increasing autonomy on the z-axis – independence from explicit human instruction – could reach surprising outputs that nonetheless cause foreseen harms.

A more extreme Row III case would be when the foreseen harm is caused through the autonomous system's 'emergent behaviour'. We saw in Chapter 1, systems that are increasingly independent from explicit human instruction are also often highly complex. Emergent behaviour is a phenomenon to which highly complex systems are subject; emergent behavioural properties of the system arise from interactions among multiple elements, including among the internal components of the machine and its local interactions in the world (Johnson, 2006a).[18] The relation of emergent behaviour to control has been described as follows:

> *"The emergent behaviour of a system cannot be reduced, more or less by definition, to the behaviour of its constituent parts. This means that the behaviour of the system as a whole cannot be completely controlled by controlling the behaviour of its parts. So, emergence and control do not go hand in hand."*
>
> (Kroes, 2016, p. 643)

[18] It should be noted that there are several alternate views of 'emergent properties', both in the context of autonomous systems and complex systems more generally. See Johnson (2006a) for a good discussion.

Though a Row IV case and not a Row III case, the emergent behaviour case in the examples given in the Introduction is the assistive robot lifting Tom.

Row IV cases

In Row IV cases – which, like Row II cases, constitute the UHC for which the *locus of moral responsibility* question seeks an answer – the output is unforeseen, and so is the consequence. In respect of the output, then, these cases are like Row III cases. Here, too, the output may be the emergent behaviour of a system. The emergent behaviour of Tom's assistive robot, which, due to unanticipated interactions between sub-optimal user commands and loud noise in the operational domain, caused an unforeseen injury to him, is a Row IV case. With emergent behaviour, the possible harms that could arise will likely be more diverse and have a wider scope than those expected to fall within accepted risk thresholds.

Another example of a Row IV case would be an output unforeseen due to an 'undetected error'. Imagine there was a bug in the assistive lifting robot software that meant it could not deal with the irregular cues from the user. It might be thought that all human errors are negligent, but it is not clear to me that this is so. Even the most conscientious designers, engineers, or operators are not perfect. Thus, I also include undetected errors as an example of a Row IV case (and for consistency include benefit of the doubt cases here, too). Cases such as these may undermine the assumptions on which risk thresholds are calculated and accepted.

To conclude, Row IV cases include unforeseen harms caused by emergent behaviour and undetected human error.

In the next Chapter, I refer back to Table 1, and specifically to Row II and Row IV cases, which denote UHC. This discussion has allowed us to see how an unforeseen harm might come about: it might be directly caused by an autonomous system in a genuine accident or as an upshot of honest human mistakes in design or deployment (Row II cases), or it might be the consequence of emergent system behaviour, or an undetected human error (Row IV cases). I have also included here 'benefit of the doubt' cases, where it is not clear whether the mistake is honest, or the lack of foresight reasonable.

2.2 The *locus of moral responsibility* question

Having clarified what I mean by the 'unforeseen harm' for which moral responsibility is sought, we may now state the *locus of moral responsibility* question explicitly: who (if anyone) is morally responsible for an unforeseen harm caused directly by an autonomous system, and on what grounds?

The answer to the question will be explicable in terms of the following three-term relation:

S is morally responsible for UHC in virtue of P

I will have answered the locus of moral responsibility question when I indicate which S are morally responsible for UHC, and in virtue of which properties of S's relation to UHC. As Talbert notes, "what really seems to matter for moral responsibility ... is that the person is related to action in a certain way" (2016, p. 1). One might object that the three-term relation is incomplete, since it lacks the term 'S is morally responsible *to A* for UHC in virtue of P.' I include this interpersonal dimension within the set of properties P, and it is particularly salient, I will argue, to questions of blameworthiness, which also concern S's relation to the addressee, whether the moral community at large or the person harmed by UHC.

An ambiguity of 'locus' is instructive. The *locus of moral responsibility* question is asking for the 'locus' in two senses. There is 'locus' in the sense of which agent (or agents, if any) is morally responsible. This sense of locus is denoted by 'S' in the three-term relation. And there is 'locus' in the sense of the grounds, or source, or properties in virtue of which an agent is morally responsible. This sense of locus is denoted by 'P' in the three-term relation. For the avoidance of doubt, in the rest of this thesis, when I speak of 'locus,' I mean it in the first sense.[19]

[19] Naturally, the two senses are related. It would be arbitrary to ascribe moral responsibility to an agent without providing grounds for doing so. On one interpretation of Strawson (2003), discussed in §2.3, it is the first sense of 'locus' that informs the second: our participant reactive attitudes towards people such as resentment, and the blaming practices that express them, are explanatorily prior to the question of whether these are fitting and deserved (see Todd, 2016 and Brink & Nelkin, 2013 for discussion). In my answer to the *locus of moral responsibility* question, where these reactive attitudes may be uncertain, I take the more objective stance of considering what would make certain human agents fitting targets of these reactive attitudes, and so use the second sense of 'locus' to inform the first. Even so, I do think that these participant reactive attitudes are compelling, and in Chapter 5 I consider how they might indicate that certain autonomous systems themselves should be included within the set of morally responsible agents.

2.3 The conception of moral responsibility

Now that the *locus of moral responsibility* question has been formulated, it is necessary to clarify the conception of moral responsibility sought. The basis of my structure for my answer to the *locus of moral responsibility* question draws on a distinction between 'responsibility as attributability' and 'responsibility as blameworthiness'.

At its most general, derivative upon its etymological root in the Latin 'respondere' (to respond, to answer to, to pledge), 'responsibility' refers to the state of being answerable for something (Lucas, 1995, p.6). But there are manifold ways of being answerable for something.[20] One distinction is between moral, causal, and legal responsibility. If one is causally responsible for something, one is simply the cause of it, or a salient causal factor in its coming about. Events and artefacts, as much as people, can be said to be causally responsible for things (Hart, 1968, p. 214). If one is legally responsible for something, one is liable for it by law, and subject to a sanction of some kind for it, ranging from compensation to incarceration. But to be morally responsible for something is to be answerable for it in a way that opens one up to evaluations and responses of moral praise and blame from other agents. Strawson influentially, in *Freedom and Resentment*, locates these as expressions of our participant reactive attitudes, which "are essentially natural human reactions to the good or ill will or indifference of others towards us, when displayed in their attitudes and actions" (2003, p. 80).

This leads me now to a central distinction I draw in my procedure of answering the *locus of moral responsibility* question, and in my framing of ascriptions of moral responsibility within a delegation framework. I signposted it earlier in this Chapter in my discussion of strict moral responsibility, but now I can be more specific. The distinction I draw upon is between what Scanlon calls 'responsibility as attributability' and 'blameworthiness' (1998, p. 248).[22] S would bear responsibility as attributability (or be 'attributionally responsible') for UHC if and only if it is attributable to S in a way which is necessary to serve as a basis for S's moral appraisal. To say that an agent bears responsibility as attributability "is to say that an agent is open, in principle, to demands of justification regarding that thing" (Smith, 2012, p. 577-8).

[20] Hart (1968, pp. 211-230) provides the classic taxonomy.

[22] The main distinction Scanlon makes is between 'responsibility as attributability' and 'substantive responsibility' – but within responsibility as attributability, Scanlon clarifies that the attribution is a *precondition* for moral appraisal, and judgements of blameworthiness are the *upshot* of moral appraisal; it is that distinction which I draw upon.

Responsibilty as attributability means that the agent is rightly called upon to answer for the thing in question.[23] Responsibility as attributability is a precondition of responsibility as blameworthiness, which would be established by the subsequent moral appraisal of S on the basis of those answers: a judgement of blameworthiness is one possible upshot of the appraisal. If S does not have a good moral justification for UHC, for which she bears responsibility as attributability, then it would be appropriate to consider S blameworthy and reasonably subject to the negative participant reactive attitudes of others.

My procedure for answering the *locus of moral responsibility* question is to draw on the delegation relation to determine the *moral responsibility as attributability* first, and then to determine whether those agents are also *blameworthy*. Whereas attribution concerns the S's relation to the harm, blameworthiness concerns S's relation and attitudes to those affected by it. When both of these questions are answered in a principled way, and specifically for UHC, so too is the *locus of moral responsibility* question. This two-stage procedure lends itself to our focus on cases of harm where it is to some degree obscure to whom UHC is even attributable, both on account of its being causally effected by an autonomous system and on account of its being an unforeseen consequence.[24]

Another distinction to be drawn is directional. Moral responsibility can be forward-looking and backward-looking. Forward-looking responsibility can mean one of two things. First, there is forward-looking or prospective moral responsibility, also called 'role responsibility', which denotes an obligation to do or to bring about certain things or positive states of affairs. Second, are the forward-looking perspectives on moral responsibility that justify ascriptions of moral responsibility on the basis of the beneficial consequences that can be obtained from them, given that punishment and reward offer motives for good future behaviour (Smart, 1961). The conception of responsibility sought by the *locus of moral responsibility* question in this thesis is not forward-looking in either of these senses.[25] Backward-looking responsibility,

---

[23] Thus understood, my approach incorporates into responsibility as attributability that which Shoemaker has argued sits within a different category: responsibility as answerability (2015; 2013, p. 157; 20).

[24] The second stage of my procedure is often called 'responsibility as accountability' (Oshana, 2004; Watson, 1996), although 'accountability' is generally taken to extend beyond evaluations of blameworthiness. Responsibility as accountability concerns being *held* responsible by others. I speak of blameworthiness rather than accountability, because my concern is what would warrant S being deemed blameworthy by others, rather than the wider practice of being held responsible by them, and the subsequent responses of the addressee, such as decreased willingness to enter into future special relations with S or to help S should she need it (Scanlon, 2013, pp. 105-106).

[25] Naturally, one can be retrospectively responsible for harms caused as a result of failing to discharge one's prospective duty, so the two are related.

in contrast to the first sense of forward-looking responsibility just given, is responsibility for events or states of affairs after they have occurred. In contrast to the second sense of forward-looking responsibility just given, as discussed above, my concern is desert-based. That is, my principal concern is whether moral responsibility for the harm would be deserved, and the reasons for which it would be deserved, not whether public ascriptions of moral responsibility would induce better future consequences. The conditions of moral responsibility examined in the next sub-section are intended to be taken as conditions for retrospective and deserved moral responsibility.

Despite my separation of attributability from blameworthiness, however, sometimes it will more be appropriate to speak of them indistinguishably. In part this is because our traditional conceptual framework for ascribing moral responsibility – the four conditions discussed in the next sub-section – has generally assumed that whether one *did* the thing at all is taken as a given (once it has been established that the agent meets certain agential criteria). Before looking at these four conditions, I should clarify that nothing in my conception of moral responsibility requires me to engage with problems of free will and free agency, and that the core argument presented in this thesis does not rely on any position in this metaphysical debate.

2.4 The four conditions of moral responsibility

We can now look in a little more detail at the four conditions of moral responsibility which, on a fairly uncontroversial account, would normally be individually necessary and jointly sufficient for a warranted ascription of moral responsibility to S. This set of conditions captures the required property P of the agent S (condition 1) and the requisite properties of S's relation to the event (conditions 2, 3, and 4) to warrant or ground a responsibility ascription.

The first condition is 'agential capacity' – that S is sufficiently able to reason, and has the appropriate psychological characteristics, to bear moral responsibility at all (Hart, 1968, p. 227). The second condition is a weak causal condition, that S was a causal factor in the causal pathway to the harm for which moral responsibility is sought. The third and fourth conditions are often called the 'Aristotelian conditions' (Coeckelbergh, 2020a, p. 2052;

Fischer & Ravizza, 1998, p. 12), because they can be traced back to Aristotle's classic account of voluntariness in Book 3 of the *Nichomachean Ethics* and the kind of actions for which we can rightly be praised and blamed: those over which we, and not another agent, exercised control; and those that were not performed in ignorance (Aristotle, 2002, 1109b30-1111b1).[28]

Drawing on a distinction made by Watson, the first two conditions – agential capacity and causal responsibility – are *exempting* conditions. Failure to meet them would render S exempt from, or ineligible for, moral responsibility.[29] The third and fourth conditions – control and knowledge – are *excusing* conditions. Failure to meet them would mean that S would be excused from moral responsibility (2004, pp. 223-4). The control condition is highly salient to determinations of responsibility as attributability, for if one is not in control of or controlling an action then it seems that one is not doing it, in any authorship sense, at all. But we can also ask whether this loss of control is blameworthy, due to a recklessness or a disregard for risk to others. Nelkin & Rickless recognise acutely that "the epistemic and control conditions are not in fact two separate conditions; rather the control required for moral responsibility itself requires a kind of awareness" (2017, p. 109). But in its own right, the knowledge condition is also salient to responsibility as attributability. It would be difficult to warrant that someone is the author of an outcome when she is not aware of what she is doing, such as alerting the prowler by turning on the light, or poisoning the inhabitants of a house by pumping the well, to use Davidson's and Anscombe's now stock examples (Davidson, 1963, p. 686; Anscombe, 1957, p. 37). Yet here, too, we can ask whether the ignorance was negligent or culpable. For this reason, it is difficult to delineate the Aristotelian conditions into neat categorisations of conditions for moral responsibility as attributability and for moral responsibility as blameworthiness.

As a methodological point, these four conditions have emerged from philosophical reflection on moral responsibility for direct actions, for one's own doings and deeds in the world. If they are to provide a framework for thinking about responsibility for harms caused by autonomous systems, they will need to be adapted to the question of extended actions, involving other actors (or autonomous systems), over periods of time (that is, diachronically and synchronically extended), which is something I consider in the next Chapter. It is not

[28] Despite calling them the Aristotelian conditions, however, it should be noted that my discussion diverges in detail from this original, classic account; it is not my intention to be anachronistic See Smiley (1992, pp. 33-37) for a discussion of this anachronistic tendency in the moral responsibility literature.
[29] To note, Watson does not discuss cause as an exempting condition; its description as such is my analysis.

clearly determinable what degree of control, and of what nature, is responsibility-grounding in these cases, and likewise what level of epistemic awareness and understanding is required. The argument that puts the most pressure on a defence of a Standard View position is the 'responsibility gap' argument from the suspension of the Aristotelian conditions - as we shall see in the next Chapter. This argument needs to show that the suspension is *sufficiently severe* to be responsibility-undermining in these cases. By focusing on the unforeseen harmful consequence, I provide a set of cases that gives this sceptical challenge the greatest chance of success.

The four conditions of moral responsibility provide a framework for articulating the 'responsibility gap' problem. I examine this responsibility gap argument in the next Chapter. It can be briefly articulated as follows. Autonomous systems, we assume, do not satisfy Condition 1. If this is true, then they would be *exempt* from moral responsibility for harms they directly cause. Human agents, by contrast, do not to satisfy Conditions 3 and 4 with respect to the consequences of the outputs of autonomous systems. If this is true, then human agents would appear to be *excused* from moral responsibility for harms the systems directly cause. As such, neither candidate loci of moral responsibility meet the requisite conditions in virtue of which an ascription of moral responsibility to them would be warranted. Therefore, a 'responsibility gap' problem seems to take hold.

2.4.1 Condition 1: Agential capacity

Agential capacity is what qualifies an agent for moral responsibility ascriptions at all. It is what makes an agent a fitting subject of moral appraisal and a fitting target of participant reactive attitudes. In the law, agential capacity is connected to one's capacity to act with *mens rea*, with intention, and to have sufficient rational capacities to be a legal subject. People with severe learning difficulties, severe mental disorders, children, and animals would all be excluded on these grounds. Mature, unimpaired adults paradigmatically meet the condition. This is not just a criterion for law. A person's 'capacity responsibility', as Hart calls it, is also a fundamental criterion of moral responsibility (Hart, 1968, p. 227).

One of the agential capacities required as a precondition for morally responsible agency, and hence for meeting the first condition, is the capacity for *moral* agency. Philosophers often take moral agency and morally responsible agency to be equivalent. In different ways, this

equivalence has been due to the influence by Locke (for whom the 'forensic term' of 'person' is defined in terms of a continuity of consciousness, which grounds ascriptions of moral responsibility), Hume (for whom moral agency involves a developed and reflective moral sensibility and capacity for sympathy) and Kant (for whom moral agency is defined with a free agent's capacity to act for the sake of the moral law) — these are rich conceptions of what it is to be a moral agent.[30] More recently, however, interest in the question of whether non-human animals or non-human agents, such as robots or corporations, can be moral agents has led to a thinning of the concept of moral agency.[31]

I follow a thin conception of moral agent, such that "the 'responsible' in 'morally responsible agent' is a non-trivial qualification" (McKenna, 2012, p. 2). That is, moral agency is a necessary but not a sufficient condition of moral responsibility. My thin conception of a moral agent is an agent capable of acting morally, where this means, on my understanding, capable of responding to other-regarding or moral reasons, which in turn implies intentional agency.[32] I take reasons to be considerations that count in favour of some action. On this thin conception, not only mature human adults, but also children, some animals, and arguably corporations and institutions, are capable of being moral agents. Though it is my working assumption that autonomous systems are not agents of this kind, as I shall discuss in Chapter 5, it is *possible* that autonomous systems *could* be.

Morally responsible agency requires a deeper capacity than being able to act morally in this way. I work on the basis that it requires the general agential capacity to meet the control and knowledge conditions with respect to one's own actions – and, specifically, to exercise reasons-responsive control over one's own actions, and to understand or be aware of what one is doing, and to have some capacity for foresight of the consequences. Moreover, a morally responsible agent recognises herself as a subject of the participant reactive attitudes of members of the moral community. Naturally, these deeper capacities can be diminished, impaired, or absent in moral agents (Hart, 1968, p. 228). Some moral agents may possess some but not all of these capacities to a qualifying degree. This has led Shoemaker (2015) to

[30] See Locke (1689, Book II, Chapter xxvii), Hume (1751), Kant (1785)

[31] For discussions of the moral agency of non-human animals, see: Clement (2013); Shapiro (2006); Sapontzis (1987). For discussions of the moral agency of corporations, see: List & Pettit (2011), Pettit (2007), May (1987), French (1979) and the discussion in Sepinwall (2016). Discussions of the moral agency of robots and autonomous systems will be considered in Chapter 5.

[32] My conception of 'moral agent' differs from McKenna, for whom moral agency requires an understanding of moral predicates (2012, p. 11)

offer an analysis of moral responsibility for 'marginal agents' by which people who have impaired faculties in some respect are not responsible in the correlatively appropriate respect, but are responsible in other respects. Psychopaths, for example, have empathic impairments which mean that they lack sufficient regard for other people to be accountable to them, on Shoemaker's account, but psychopaths may still bear responsibility as attributability (2015, p. 173).

I assume that, all things being equal, high-level and low-level human agents *do* meet the agential capacity condition. As I discuss below (see §4.5 and §6.2.2), I assume this is true of them as individuals acting alone, and also, as will more likely be the case in real-world relations of high-level and low-level human agents to autonomous systems, in both loose assocations and as members of 'plural agents' By plural agents, I mean groups of individuals pursuing a joint project, and working together to achieve a shared goal (Ludwig, 2017; Bratman, 2014; Gilbert, 2014; Bratman, 1993). I have in mind such groups as a design and product team within a manufacturer, or a team of clinicians on an ICU ward. These individuals broadly share their intentions (to create a particular product; to deliver the best possible medical care), which have a distinctive character in virtue of their plural context. Each individual member of a plural agent bears moral responsibility in her own right for her decisions made as part of the plural group, in pursuit of the shared goal or project.

This might lead to the observation that I have excluded another kind of group agent from the picture. Following List and Pettit, we might call it the 'single group agent' to distinguish it from the plural agent (List and Pettit, 2011). Whereas plural agents are reducible to their individual group members, single group agents, such as corporations and institutions, transcend or supervene upon their individual membership. Technology corporations, institutions, and public regulatory bodies are involved upstream in the development and also downstream in the deployment of autonomous systems. So how do such agents fit into the picture? Do they meet the first agential condition of moral responsibility? And, if so, what would the implications be for a Standard View position that only human agents are morally responsible for harms caused by autonomous systems?

I shall take the second of these questions first, since it leads me to sharpen my articulation of the Standard View position. Single group agents are not themselves human agents. As such, by definition, I exclude single group agents from the Standard View. Any account,

therefore, that successfully maintains that single groups agents like technology corporations are even to some degree morally responsible for the harms would be an account that constitutes a rebuttal of the Standard View as I have defined it, since this was the position that human agents are exclusively responsible for harms the systems cause.

We therefore need to address directly whether single group agents of this kind can be morally responsible agents. List and Pettit, developing a view first put forward by Pettit (2007, p. 175), offer an argument in defence of ascriptions of moral responsibility to single group agents. They argue that group agents can meet the following conditions of responsibility: *normative significance* - they can face a normatively significant choice, involving the possibility of doing something good or bad, or right or wrong; *judgmental capacity* - they can have the understanding and access to evidence required for making normative judgements about those options; *relevant control* - and they can have the sort of control required for choosing between those options (List and Pettit, 2011, p. 155).

Setting aside the first of these conditions, the second two conditions proposed by List and Pettit are broadly consistent with the knowledge and control conditions presented in this Chapter. I have said that the capacity to meet these conditions furnishes the rich agential capacity that is a prerequisite for being a morally responsible agent in general. Let me state here my position. I think it is at least plausible that single group agents can act morally, in response to moral reasons. A national health regulator, for example, can approve a vaccine in response to reasons such as the health and well-being of citizens. This capacity, on my account, confers thin moral agency. Although the more controversial dimension of List and Pettit's argument is that single group agents can have the requisite control over their actions to ground a responsibility ascription, I do not rebut this claim. In §2.4.3, I elaborate on what kind of control this would be: a reasons-responsive control which issues from the deliberative mechanisms of the agent. In the case of single groups agents, it seems plausible to say that certain organisational structures and procedures, such as the votes of its board or authorized subgroups, constitute these deliberative mechanisms.

It is, however, I think, conceptually untenable to think of the single group agent as an entity that can *understand* what it is doing and hence meet the knowledge condition. List and Pettit flesh out their version of the knowledge condition thus: "the group agent must be able to form judgements on propositions bearing on the relative value of the options it faces –

otherwise it will lack normative understanding – and it must be able to access the evidence on related matters" (2011, p. 158). Forming a judgement on a proposition is a plausible description of what happens when, for example, an authorised subgroup votes on an agenda item in a key meeting; the subsequent judgement becomes a single group judgement. But the single group agent's forming of these judgements does not *yield* understanding, as the authors claim. Rather, the understanding *informs* the judgements. The individuals in the sub-group understand what it is at stake, understand the relative values of the different options, and the significance of the group level action to which the judgement will foreseeably give rise. This understanding is possessed by the individuals and not by the single group agent; to say that the single group agent understands is merely elliptical for saying that these individuals within the group do. On grounds of failing to meet the understanding dimension of the knowledge condition, therefore, it is my position that single group agents do not meet the agential capacity condition of moral responsibility. As such, I endorse a form of methodological individualism, such that moral responsibility (though not, on my view, thin moral agency) is properly ascribed to individuals and not to single group agents.

The common consensus is that autonomous systems fall short of the requisite agential capacity to bear moral responsibility, although the arguments that they fall short in this way may be grounded in different proposed necessary conditions of responsibility that the systems fail to meet, such as the systems' lack of consciousness (Himma, 2009) or their incapacity to suffer from punishment (Sparrow, 2007). I interrogate this common consensus in Chapter 5.

2.4.2 Condition 2: A degree of causal responsibility

Moral responsibility entails causal responsibility, where 'causal responsibility' means being a cause of the action, omission, or consequence. One cannot even be in the ballpark for a responsibility ascription if one has not even been a causal factor in the outcome. But they are not directly assimilated (Thompson, 1980, p. 909). Degrees of moral responsibility do not rise and fall with degrees of causal weight. The responsible agent need not be the primary cause of the outcome, simply a salient causal factor in its coming about. I therefore modify 'causal responsibility' with 'a degree of', such that this causal condition of moral responsibility becomes a very weak one. Being the primary cause is neither sufficient nor necessary for moral responsibility. Infants and animals, for example, can be the primary cause of

occurrences for which they are not morally responsible, because they are generally thought to be exempt from moral responsibility ascriptions. Thus, primary causation is not sufficient for moral responsibility. Even agents who are not generally exempt from responsibility ascriptions may be the primary cause of things for which they are not responsible: when, for example, they cause things involuntarily. In these cases, too, primary causation is not sufficient for moral responsibility. And agents may be morally responsible for outcomes in which they were not the primary cause: parents can be morally responsible for the consequences of their young child's mischief. *Primary* causation is not therefore even necessary for moral responsibility. But being a causal factor, which I shall take to mean the same thing as bearing a degree of causal responsibility, is necessary for moral responsibility. If an agent plays no causal role whatsoever in the production of an outcome, she would be exempt from responsibility as attributability. I assume, for the most part, that all the candidate loci of moral responsibility for UHC – both the humans and the autonomous systems – meet the second condition, so weakly construed. Doing so is what makes them candidates in the first place.

2.4.3 Condition 3: Control

No one ever has complete control over an action. Each of our actions will be contingent upon some things that are beyond our control, such as the fact that we were born at all (Zimmerman, 2006, p. 591). A condition of complete control sets the bar too high for moral responsibility. What follows, then, should be understood as within the realm of agency that is partial control. Within this realm, there are two further distinctions that are salient to our inquiry. These will be applied to the case of high-level and low-level human agents in the next Chapter, when I consider whether they are vulnerable to the objection from the responsibility gap position that they lack sufficient control over the system to warrant moral responsibility ascriptions for its behaviour (and its consequences). The first distinction is between direct and indirect control. The second distinction is between regulative and guidance control.

My understanding of direct control connects to the standard stories of action. I directly control the intentional raising of my arm, the writing of these words, my drinking of this tea, and so on. This is contrasted with indirect control, which is my agency exercised through an

intervening mechanism, such as via a lifting device for my paralysed arm, or some voice recognition technology that transposes my spoken words to text. Mele puts it thus,

> *"The following certainly seems to be a plausible requirement on an agent's exercising direct control over X: If S exercises direct control over X, then S does not exercise control over X only by exercising control over something else (or, more precisely, something that does not include X)."*
>
> (Mele, 2017, p 280)

It is clear that human control over autonomous systems would not be direct. Indeed, this is analytic, given my conceptualisation of autonomous systems in terms of their degrees of independence form direct human control on three dimensions. But there is no reason, *prima facie*, why indirect control alone should cause us to question whether the indirect controller bears responsibility for harms she thereby causes. This would rule out vast numbers of cases in which we act through mechanisms and devices, whether technological or institutional. The challenge of a responsibility gap starts to become serious when there are constraints on even this indirect control, as I shall explore in the next Chapter, and which is illustrated in many cases of unforeseen harm.

The notion of meeting the control condition was originally couched by Aristotle as being exemplified in voluntary actions, in which the origin of the action is in the agent herself. In modern parlance, what happens is 'up to' the agent. By contrast, involuntary behaviour, over which we do not exercise control, has a cause external to the agent, such as being carried away by the wind, or she acts under force of others (Aristotle, 2002, 1110a1). Here, what happens is not up to the agent. This central insight has, throughout the history of the philosophy of moral responsibility, given rise to the idea that the responsibility-grounding control requires the ability to do otherwise. This long-standing principle – the Principal of Alternative Possibilities – was influentially challenged by Frankfurt (1971, 1969). This brings us to the second distinction, between regulative and guidance control.

'Regulative control', which is a term used by Fischer & Ravizza (1998, p. 31), corresponds to the ability to do otherwise. Frankfurt-type cases demonstrate that regulative control is not necessary for a warranted ascription of moral responsibility. They involve the presence of what Fischer & Ravizza call "a counterfactual intervener" (1998, p. 29). The cases have the following structure. An agent makes a decision at t1 to do something at t2. Unbeknownst

to her, at around t1, an outside party (a counterfactual intervener) has planted a device in her brain that will forcibly cause her to do that thing at t2 – the examples usually involve shootings and assassinations – in the event that she wavers on enacting the decision she made at t1. Even so, at t2 the agent does not in fact waver but acts on her decision without the interference of this device. The agent did as she chose, but could not have done otherwise. Intuitively, she is still morally responsible for what she did in fact do. Regulative control is therefore not always necessary for an ascription of moral responsibility. Even so, it is sufficient.

What obtains in the Frankfurt cases, and which Fischer & Ravizza in particular think explains when and why agents are *still* morally responsible in such cases, is that the action at t2 still issues from the agent's processes of reasoning and practical deliberation, which they call her 'decisional mechanism'. A similar notion has been expressed by Dennett: "A controls B if and only if the relation between A and B is such that A can drive B into whichever of B's normal range of states A wants B to be in" (Dennett, 1984, p. 52). If S has actual sequence control over what happens, and this control is responsive to her reasons, including her moral reasons, then ascriptions of moral responsibility are warranted irrespective of whether some possible intervention could have caused her to do otherwise. Actual sequence reasons-responsiveness is generally taken to be sufficient for warranted ascriptions of moral responsibility. This actual sequence form of rational control Fischer & Ravizza call 'guidance control' (1998, p. 31). Fischer & Ravizza's account is a particularly fertile account of responsibility-grounding control for the context of autonomous systems. It can be adapted to the technological case (Santoni de Sio & van den Hoven, 2018, p. 6). In addition to its extendibility to cases involving technological devices and artefacts, Fischer & Ravizza's analysis of responsibility is a common cornerstone in discussions of human responsibility in respect of autonomous systems (Coeckelbergh, 2020a; Di Nucci, 2020; Santoni de Sio & van den Hoven, 2018; Matthias, 2004). Moreover, they explicitly and quite helpfully provide an analysis of responsibility for a consequence, which is the logical object of our inquiry – UHC (Fischer & Ravizza, 1998, pp. 91 - 122). Their account therefore provides a good anchor for consideration of the causal pathway from the second to the third column of Table 1, from the output to the harm.

Di Nucci (2020, p. 40) and Di Nucci & Santoni de Sio (2014, p. 6) make a distinction between human controlling and overall control. This forms a central core of Di Nucci's argument

that loss of human controlling does not entail loss of human responsibility, since *overall* human control trumps it (Di Nucci, 2020). For Di Nucci, 'controlling' maps onto guidance control, and 'overall control' (or being 'in control') maps onto regulative control (Di Nucci, 2020, p. 49). Di Nucci also calls controlling, or guidance control, 'direct control', because it denotes real-time, conscious, and continuous control over the thing or event. But since I think there is scope for what we might call indirect guidance control, or indirect controlling, on the alternative terminology – that is, reasons-responsiveness exercised via an intervening mechanism, such as a computational system – I restrict my use of 'direct' to mean 'not done by exercising something else'.

In the next Chapter, I consider whether high-level or low-level human agents meet the necessary control condition for moral responsibility: whether they have regulative control or guidance control. Regulative control would obtain if they were able to intervene in the operation of the system, whether directly, through the resumption of operational control, as would have occurred if Jacinta had taken the wheel, or indirectly, through intervening mechanisms in the system, as occurred to some extent when Myrtle went into a minimum risk procedure. My account of guidance control is structured in accordance with Fischer & Ravizza. It simply requires extending to the autonomous system's output rather than to a person's direct action. The conditions for guidance control can be stated briefly here. First, an ownership requirement, such that the output issues from the agent's 'decisional mechanism' (1998, p. 39, p. 62 & p. 241). Second, a receptiveness requirement, such that the output responds to the agent's relevant reasons and considerations, which include her moral reasons (1998, pp. 69-72). Third, a reactivity requirement, such that the output is a concrete implementation of this response to the agent's reasons (1998, pp. 73-76). In addition, for guidance control over the consequence and not just the output, the consequence should be suitably sensitive to the output; that is, it would be reasonable to expect it to follow from the output, and the harm comes about primarily because of the output (1998, p. 101). This latter requirement is to cover the fact that if someone were to injure someone by deflecting a threat caused by a third party, it would be unfair to ascribe moral responsibility to them for the injury. Take, for example, the case of Morgan and the homework bot. Morgan's psychological distress *was* suitably sensitive to the robot's informational outputs in this way. Where these sets of conditions are met, respectively, at both stages (both guidance control over the output, and guidance control over the consequence), S would meet the guidance condition for moral responsibility.

2.4.4 Condition 4: Knowledge

As we have seen, on the Aristotelian conditions, it is not only loss of control that renders one excused from being open to moral appraisal for any harms, it is also ignorance of the actions one performs and their likely consequences (Aristotle, 2002, 1110b20 – 1111a20). Very often, ignorance excuses. "I didn't know", when truly uttered, is related to "I didn't mean to" – both of which generally would cause us to reassess our reactive attitudes, and take into account this further, mitigatory evidence. As I have delineated the specific harm for which responsibility is sought as UHC – the genuinely unforeseen harmful consequence – it is necessarily true that these are precisely the cases in which the knowledge condition is not met by the low-level and high-level agents.

It is striking that in the philosophical discussion of the knowledge condition, which is less extensive than the discussion of the control condition, with its long historical connection to the question of free will, is far more focused on blameworthiness than on attribution (Robichaud & Wieland, 2017; Smith, 1983). This may in part be down to the overall emphasis of this conceptual framework on individual personal actions, to which direct epistemic access or immediate "knowledge without observation" are taken as given (Anscombe, 1957, p. 13). Even so, the knowledge condition naturally lends itself to further considerations such as whether the fact that it was unforeseen was avoidable, and whether failing to foresee it in and of itself is blameworthy – whether it is derivative upon an earlier negligent decision or action, where due standards of care and concern should have been met. Since I stipulated earlier in this Chapter that UHC is the reasonably unforeseen consequence, my discussion of whether high-level or low-level agents meet the knowledge condition for moral responsibility rules this out by fiat in the next Chapter.

It might seem that my focus on these rarefied cases, in which human actors fail to meet the knowledge condition by definition, is misguided. These cases, which are putatively harder for the Standard View, are harder, the objection might go, because they'd be exculpatory in all cases. What does this specific focus on the non-negligently unforeseen harmful consequence add to determinations of responsibility for autonomous systems? But it is still instructive to determine responsibility as *attributability* in these cases. Even if they are cases in which agents are not *blameworthy*, we should still have a framework for establishing which agents rightly warrant the responsibility attribution in the first place. In any case, I argue in

Chapter 4 that non-negligent lack of foresight of UHC does not *alone* excuse an agent from blame (see §4.4). To be clear, the methodological benefit of answering this harder question of responsibility for unforeseen harmful consequences is that the response to it yields a principle that traces responsibility as attributability for *all* cases of harm, and not just for foreseen harms, the risk of which had already been accepted. Thus, the focus on the rarefied cases allows us to defend the 'completely' element of the Standard View – that humans are not just exclusively but completely responsible for harms caused by autonomous systems; that is, that there are no gaps and no cases fall outside of human responsibility. Moreover, in the next Chapter, it is interesting to see the plausible diverse ways such unforeseen harm may come about. And this shows us that the apparently rarefied cases are not, perhaps, so rarefied after all, and that we best have an answer to them.

One final point on the Aristotelian conditions. They are subject to a well-known catch. This is the caveat that if the agent is responsible *for* the loss of control or ignorance, the agent cannot be excused *by* the loss of control or ignorance. This will be a feature of my defence of a Standard View position in Chapter 4, though with some refinement. The central derivative structure of this catch to the Aristotelian conditions is shared by tracing views of moral responsibility, that one's moral responsibility for an outcome depends not one's control or knowledge at the time of the lapse, but on one's control or knowledge at some relevant prior time, and on the appropriate connection of facts between those two points in time (Nelkin & Rickless, 2017; Fischer & Tognazzini, 2009).

Summary of Chapter:

A defence of the Standard View position will have to identify which humans S are morally responsible for UHC and provide the reasons or grounds, furnished by P, that justify this ascription. Doing so will answer the 'locus of moral responsibility' question as I have formulated in this Chapter. Typically, P are the four individually necessary, jointly sufficient conditions of moral responsibility: agential capacity; cause; control; and knowledge. In the next Chapter, I consider the objection that the last two of these are not met by human agents with respect to UHC, and that these cases, at least, therefore pose a responsibility gap.

CHAPTER 3
THE SCEPTICAL CHALLENGE: THE 'RESPONSIBILITY GAP' PROBLEM

Having set out the 'locus of moral responsibility' problem, I consider in this Chapter the main objection to being able to ascribe moral responsibility to human agents for harms caused by an autonomous system. This is the objection that the Aristotelian conditions, which are necessary excusing conditions, are not met adequately by humans for it to be fair to ascribe moral responsibility to them for these outcomes. It is a common premise in 'responsibility gap' arguments. This Chapter examines whether that premise is true.

3.1 Sceptical challenges to the Standard View

This thesis defends the Standard View position that human agents are completely and exclusively morally responsible for the immediate harmful consequences of autonomous systems. In the previous Chapter, I set out the four traditional conditions for moral responsibility. Sceptical challenges to the Standard View target one or more of these conditions and posit that, when it comes to the outcomes of what autonomous systems do, human agents fail to meet them. In this Chapter, I first consider a challenge to the second condition (cause), and then I consider in detail the challenge to the third and fourth conditions, or the Aristotelian conditions (control and knowledge). The responsibility gap position, first articulated by Matthias (2004), is generally couched in terms of the failure of human agents to meet the Aristotelian conditions with respect to the behaviour of an autonomous system.[33] When I speak of the 'suspension of the Aristotelian conditions', I mean this failure to meet them. Responsibility gap arguments take as their reference point systems with high degrees of autonomy on what I have called the x-axis (the system's operational independence from direct human interference) and the z-axis (the system's independence from explicit human instruction). I argue that, while the claim of the suspension of the Aristotelian conditions is in danger of being overblown, it is sometimes true, and the tricky cases of unforeseen harm fall into this category.

[33] To note, this is not the only way in which the responsibility gap argument can be presented. Himmelreich presents it thus: "a situation gives rise to a responsibility gap if and only if (1) a merely minimal agent does x, such that (2) no one is responsible for x; but (3) had x been the action of a human person, then this person would be responsible for x." (2019, p. 735). My articulation of the problem in §2.4 provides a reason for supporting (2), namely that none of the plausible candidate meet the necessary and jointly sufficient conditions for moral responsibility.

It should be noted that my separation of moral responsibility as attributability from moral responsibility as blameworthiness is not standard in the responsibility gap literature.[34] The problem is usually taken to be, sometimes explicitly, more often by implication, an accountability problem, which includes questions of blame, sanction, and punishment (e.g. Habli, Lawton & Porter, 2020; Sparrow, 2007). Given this generality, and also the inherent generality of the four conditions – as discussed in the previous Chapter – to some extent, this Chapter will also be more general, before I return to the distinction between attributability and blameworthiness more sharply.

3.2 The problem of many hands

Before looking at the main claim, that the Aristotelian conditions are not adequately met for responsibility, I will briefly consider another common sceptical challenge to warranted ascriptions of moral responsibility in complex cases like these. It is the Problem of Many Hands. Whereas the challenge that the Aristotelian conditions are suspended concerns the third and fourth conditions of moral responsibility, this problem appears to present a challenge for fulfilment of the second condition: cause. The original statement is due to Thompson (1980), who formulated it to articulate a concern about the personal moral responsibility of public officials. It has since been adapted to numerous other spheres in which the loci of responsibility are apparently occluded by the number of decision-makers involved, including in the design of computational systems, as well as big business, government, and the military (Van de Poel, Royakkers & Zwart, 2015; Nissenbaum, 1996).

The Problem of Many Hands is the problem that many outcomes "are the product of the actions of many different people whose individual contributions may not be identifiable at all, and certainly cannot be distinguished significantly from other people's contributions" (Thompson, 1980, p. 907). Autonomous systems clearly meet the requirements of that central predicament. They are the product of the actions of many different decision-makers. Design and product teams sit within large corporations or research institutions. Regulations that inform, guide, and constrain this development are determined by networks of public officials, in collaboration with academics, and industry representatives. And on the buy side,

[34] One exception to this is Simpson & Müller (2016, p.6), in their brief drawing of a distinction between a 'responsibility gap' and a 'blameworthiness gap'.

while autonomous systems will sometimes be bought and deployed by lone individuals, they will often be procured by large public sector organisations or companies. The Problem of Many Hands is that it is not possible to trace back exact causal contributions of each of these very many causally involved actors.

One interpretation of the Problem of Many Hands is that it is an instance of epistemic opacity. We may refer to this interpretation as 'Many Hands as Epistemic Opacity', or MHEO. What I mean by 'epistemic opacity' here is that it is inherently difficult to *know* the causal contribution of each individual agent. This is the interpretation given by Nissenbaum, who argues in response that we should not confuse *obscurity* of responsibility with *absence* of responsibility (1996, p. 32). Even so, it must be recognised that, plausibly, a soft feasibility constraint applies here, such that even if not logically impossible, it would be so practically difficult to know each agent's causal contribution in such cases that this could trigger the ought-implies-can principle and it could not be a reasonable expectation or duty upon the responsibility-seeking party.[35] Even though our concern is retrospective responsibility for a harm and not the moral duties of observers seeking to locate that responsibility, this problem does indicate that epistemic transparency with respect to each agent's causal contribution would not be an implementable requirement for an ascription of moral responsibility. However, knowledge of each individual agent's causal contribution is not in any case *necessary* for such an ascription. This is because the precise causal contribution made by each individual agent is in fact not something that tracks moral responsibility anyway.

This feature of the relation between causal and moral responsibility becomes relevant also to the second interpretation of the Problem of Many Hands, which we may refer to as 'Many Hands as Causal Opacity', or MHCO. This is the interpretation that the outcome for which responsibility is sought is an instance of actions that are causally inextricable, and in which the individual causal contributions to the outcome are so weak that the outcome could only be causally attributable to the collective and not to any individuals within it.[36] This is one possible interpretation of the suggestion made by Bovens that the Problem of Many Hands

[35] By 'soft feasibility constraint' I mean the kind of constraint that is discussed in particular in the political philosophy literature to denote the sort of constraint that makes it infeasible to implement an action because doing so would be extremely difficult or costly, perhaps due to empirical and psychological factors. A hard feasibility constraint, by contrast, renders an action impossible to perform (see, for example, Gilabert and Lawford-Smith, 2010, pp. 813-816).

[36] I use the term 'collective' as an agnostic term that covers both plural agents and single group agents as discussed in Chapter 2 (§2.4.1).

occurs if a collective is responsible for an outcome but none of the individuals in the collective are responsible (Bovens, 1998, p. 47, cited in Van de Poel, Royakkers & Zwart, 2015, p. 51). But to recall, the causal condition as I set it out in the previous Chapter (see §2.4.2) was quite explicitly a weak one. It was the weak, exempting condition that, without any causal role, or without being at least causal factor in the outcome, an agent would not even qualify as a candidate for moral responsibility. There was no suggestion that the causal condition needed to be met strongly in order to be met adequately. The claim that the Problem of Many Hands entails that only the collective and not any individuals could be ascribed moral responsibility was precisely the interpretation that Thompson in his original statement of the problem sought to refute. Thompson's refutation was to argue that causal responsibility is not exactly assimilated to moral responsibility (Thompson, 1980, p. 909). Individuals are not necessarily morally responsible to the degree that their causal influence in bringing about the outcome was weighty or strong. Other things matter, too, such as the degree and nature of their control over the relevant events, or their capacity to change internal procedures within the operation of the collective. Even if causal attribution can only rest with a group or collective, this does not mean that none of the individuals are morally responsible for the outcome.

I have introduced the Problem of Many Hands as one that targets the causal condition of moral responsibility. But let us now consider an interpretation of the Problem of Many Hands that does not depend on considerations of the *causal* attribution of the outcome but, rather, highlights a deeper problem concerning the *moral* attribution of the outcome. We may refer to this interpretation as 'Many Hands as Moral Opacity', or MHMO. This is the more difficult challenge for the Standard View position. It is the interpretation that ascribes moral responsibility to the collective because none of the individuals within the collective sufficiently meet the other conditions of responsibility, such as control and knowledge, with respect to the undesirable outcome. This is the interpretation given by van de Poel, who puts it thus: "A collective might know things that none of the individuals in the collective knows, and might be free to do things that none of the individuals in the collective is free to do" (van de Poel, Royakkers & Zwart, 2015, p. 89).

This now prompts us to consider the ontology of the collective being referred to. If the collective is an occasional collection of individuals that jointly caused a harm, or if the collective is what I have called a plural agent of individuals working together with shared

goals and shared intentions, then the collective is still reducible to individuals. Those individuals might not meet the requisite necessary conditions of control or knowledge with respect to the outcome to each be *fully* morally responsible for that outcome, but each agent can still bear their *own share* of moral responsibility for the part they played in bringing that outcome about. This does present some difficulties around shared moral responsibility, such as the worry that responsibility might be too scattered or distributed across numerous agents for it to be satisfactory. Given that my overall argument in this thesis is one of shared moral responsibility, I address such problems in Chapter 6 (see §6.2.2). But it does not present a problem for the Standard View that human agents are completely and exclusively morally responsible in cases of harm caused by autonomous systems. The Standard View does not mandate against *shared* human moral responsibility.

If, however, the collective is what I have called a single group agent, which transcends or supervenes upon its individual members, then it appears we have a more serious problem for the Standard View, which has ruled out single group agents by definition because they are not *human* agents. But as I argued in §2.4.1., I do not think it is conceptually apt to attribute moral responsibility to such agents because it does not make sense to say that these groups (and not their key members) understand the basis of their normative judgements and hence that they really meet the knowledge condition. Thus MHMO, where the collective in question is a single group agent, fails because such entities do not meet the agential capacity condition of moral responsibility, and it does not make sense to say that such a collective 'knows' things that none of its individuals know. The Problem of Many Hands does not therefore pose an insurmountable challenge to the Standard View.

3.3 The Aristotelian conditions

Let us now consider the suspension of the Aristotelian conditions from a different angle, as it is generally articulated in the context of autonomous systems – as the main premise of 'responsibility gap' arguments. In the following two sections, I consider the degree to which this putative suspension of the conditions might be true.

In its original articulation by Matthias (2004), the 'responsibility gap' problem is that:

> *"... there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to assume the responsibility for them. These cases constitute what we will call the responsibility gap"*
>
> (Matthias, 2004, p. 177)

Matthias' explicit characterisation of the problem is in terms of the control condition: nobody has enough control over the machine's action to assume responsibility for them. But this suspension of control is partly explained in terms of the inscrutability of learning systems and genetic programming methods (2004, p. 182). Matthias' analysis of the problem depends upon a reasonably high level of machine autonomy, of systems that continue to learn and to adapt in the operating environment, updating their rules independently of direct human intervention or oversight.

The next major statement of the responsibility gap is by Sparrow (2007). His account is concerned specifically with the loci of responsibility for harms caused by autonomous weapons, but the argument generalises. Sparrow goes further than Matthias in the properties he ascribes to the systems to pose the problem:

> *"... the actions of these machines will be based on reasons, but these reasons will be responsive to the internal states – 'desires', 'beliefs' and 'values' – of the system itself. Moreover, these systems will have significant capacity to form and revise these beliefs themselves. They will even have the ability to learn from experience."*
>
> (Sparrow, 2007, p. 65)

Sparrow is officially agnostic on the prospect of Strong AI, stating simply that the more autonomous the systems are, the less morally accountable the programmers and the users can be held to be (2007, p. 66). But in his imagining of a scenario in which the responsibility gap might open up in practice – the bombing of enemy soldiers who have clearly indicated their desire to surrender, which would be a war crime if committed by a human – Sparrow depicts an airborne, AI-enabled AWS that "with full knowledge of the situation and the likely consequences ... [and having] reasons for what it did; perhaps it killed them because it calculated that the military costs of watching over them and keeping them prisoner were too high ..." bombs the soldiers "deliberately" (2007, p. 66). If these predicates are to be taken

literally, this is a system with sophisticated intentionality, and which qualifies for *mens rea*. This is an overstatement of the problem, I think. The next generation of intelligent robots that Sparrow depicts would seemingly be as responsible as humans in the same role.

Sparrow identifies the 'responsibility gap' by ruling out each of the candidate loci of moral responsibility: the machine itself; the programmer; and the military commander. He argues that a crucial appropriateness condition for blame or punishment is not met by an autonomous system, namely that it is inconceivable that it could suffer, and so it would be senseless to hold the machine accountable (2007, p. 70). It is far from clear why a system with the intentional states Sparrow imputes to it could not suffer, in some sense appropriate to its kind, nor why this would be essential in any case to a legitimate ascription of moral responsibility, which precedes questions of punishment. Sparrow rules out the moral responsibility of the non-negligent programmer and the non-negligent military commander on the grounds of failing to meet the control condition. His explanation of this is grounded in the fact that the system's programming under-determines its behaviour, which excuses the programmer, and the order given to the system likewise under-determines its behaviour, which excuses the military commander. There is a hard-to-reconcile tension between Sparrow's acknowledgement that "the risk that it may go awry is accepted when the decision is made to send it into action" and the claim that the relevant risk-takers would not be accountable (2007, p. 70). Objections to Sparrow have focused on this, that even if there is not control over the harm, there is control over the risk of the harm, which makes the military commander fairly open to moral appraisal for harms that occur (Himmelreich, 2019, p. 735). My argument is similar in many respects, though not in all details, to these responses, as I will show in the next Chapter.

The final statement of the 'responsibility gap' literature I survey comes from Hakli & Mäkelä (2019). This is also articulated in terms of the Aristotelian conditions:

> *"When humans create and employ robots and algorithms that replace human beings in decision making, they lose control and also give up knowledge of what kind of decisions are made, because the decisions are based on Artificial Intelligence and machine learning techniques, whose behaviour can sometimes be unpredictable. Even the programmers themselves are often surprised how their programs behave after the learning phase. Often, humans do not even know what kinds of actions will be performed and what their consequences will be. It thus seems that both traditional conceptions*

> *of responsibility – control and knowledge – are out of the reach of human beings, including the developers, builders, controllers, owners, and users of those machines."*
>
> (Hakli & Mäkelä, 2019, p. 259)

This provides a good springboard for the discussion in the next two sub-sections. In what follows, I look in more detail what is meant by *"lose control"* and if this is in fact true, and what is meant by *"give up knowledge of what kind of decisions are made"* and if this is, in fact, true.[37]

Sparrow is right in his central insight, I think, that autonomous systems such as these sit in "an uneasy conceptual space" (2007, p. 65). As Haugeland notes about AI more generally, cases like these are interesting because they "pull us both ways" (1998, p. 292). It is precisely in virtue of their contested agential and intentional properties that a responsibility gap problem gains traction:

> *"fuzzy at both the upper and lower ends, in which entities are sufficiently complex, and possess internal states that function as ends, such that their actions can no longer be attributed to those who set them in motion, but where they are not sufficiently well-formed moral agents to be morally responsible"*
>
> (Sparrow, 2007, p. 74).

This claim – that an autonomous system's actions (or outputs, or their consequences) cannot even "be attributed to those who set them in motion" is what is at issue, and the delegation framework provides a robust response.

3.3.1 Control Condition

In the previous Chapter, I made two distinctions. The first was between direct and indirect control.

[37] Hakli & Makela (2019), Sparrow (2007), and Matthias (2004) are not the only statements of the responsibility gap in terms of the Aristotelian conditions. Habli, Lawton & Porter (2020) and Burton *et al.* (2019) also discuss the Gap in terms of these conditions. I was the author of the relevant content on the Aristotelian conditions in these papers. I do not repeat those arguments here, since this thesis marks a substantial development of my thinking. Coecklebergh (2020a; 2020b) also couches the problem in terms of the Aristotelian conditions.

The suspension of direct human control over an autonomous system is analytic, given this thesis's conceptualisation of machine autonomy, as independence from *direct* human control on three dimensions.[38] Loss of direct control is a function of increasing autonomy on the x-axis, the y-axis, and the z-axis. The suspension of direct human control is also necessary, or inevitable, because the system is carrying out the task and the human agent is not doing so personally. In and of itself, however, loss of direct control is not problematic for ascriptions of moral responsibility. There is no reason why *indirect* control cannot ground moral responsibility. The fact that high-level human agents may indirectly control a system through fail-safe mechanisms, for example, would not mean, in and of itself, that these humans cannot be responsible for the consequences of system behaviour that relies on those mechanisms.

So, if there is any human control to be exercised over the behaviour of an autonomous system, it must be indirect. This claim would only *not* stand in cases of direct operational intervention, where a human operator takes over physical control of the system – as would have occurred, for example, if Jacinta had responded to the transition demand more quickly and taken the wheel. But at the point of her doing so, the system would no longer be operating autonomously, in the negative sense of machine autonomy that I have assumed; it would no longer be in autonomous mode at that time. The issue of relevance to the *locus of moral responsibility* question, then, is whether there is a suspension of *indirect* human control over the autonomous system, its predictions, classifications and other outputs and the actions it takes, and the consequences that these cause in the world.

As stated in the previous Chapter, I consider this question through the lens of the second distinction, between what Fisher & Ravizza call 'regulative control' and 'guidance control' (1998, p. 31). With respect to one's own personal action, the possession of regulative control corresponds to the ability to do otherwise. It means that the agent freely performed the action, and she could have freely not performed the action. The possession of guidance control corresponds to actual sequence control. It means that agent freely performed the action, even if she could *not* freely have not performed the action. Fischer & Ravizza argue that guidance control is sufficient for moral responsibility. That is, an agent is responsible

[38] To clarify, the z-axis – which denotes independence from explicit human instruction – still falls within the notion of direct control on my account, but specifically to direct rational control. When systems have high degrees of autonomy on the z-axis, the outputs they reach and the actions they take are achieved through implicit instruction, which is effected via intervening mechanisms, such as the combination of meta-algorithms and training data.

for actions that are responsive to her reasons, even if someone could have (but does not) make those actions non-responsive to her reasons.

In this sub-section, I adapt that analysis to the case of human control over an autonomous system, with a particular emphasis on human control over the unforeseen consequences of an autonomous system's behaviour. Recall that, on Fischer & Ravizza's analysis, guidance control over a consequence involves a two-stage reasons-responsive sequence. The first stage is guidance control over the action (which I call the 'output' to avoid confusion with human decisions and action). This corresponds to the second column of Table 1. The second stage is guidance control over the pathway from output to consequence. This corresponds to the third column of Table 1. Our particular focus is whether the high-level or low-level human agents in the first column of the Table possess sufficient – even if indirect – guidance control over the harm at column 3 to be said to meet the control condition for moral responsibility. For ease of reference, I repeat that Table below, and include the two-stage sequence.

Table 1(b)

	Human high-level decision to develop or human low-level decision to deploy the autonomous system	The autonomous system's output	The harm caused by the autonomous system's output
		= First stage of guidance control sequence	*= Second stage of guidance control sequence*
I	Voluntary	Foreseen	Foreseen (FHC)
II	Voluntary	Foreseen	Unforeseen (UHC)
III	Voluntary	Unforeseen	Foreseen (FHC)
IV	Voluntary	Unforeseen	Unforeseen (UHC)

I now consider systematically the respective control exercised by low-level human agents and high-level human agents over the causing of the harm, UHC. My procedure below is as follows. First, I look at how far low-level human agents meet the control condition. I look at whether their guidance control is suspended over first stage of the sequence, over the output that causes the harm. I then make a similar examination for the second stage of the sequence – that is, low-level human agents' guidance control over the pathway from output to harm. I then look at whether they have regulative control over any aspect of the causing

of the harm. Second, I look at how far high-level human agents meet the control condition, following the same structure: guidance control over the first and second stages of the sequence; and then regulative control. Our purpose is to see whether these are *severe enough suspensions to control* that would give credence to the common premise of responsibility gap arguments that no one has enough control over the system's behaviour to be fairly ascribed moral responsibility for it.

The first stage concerns the 'output'. Following Fischer & Ravizza, guidance control has an ownership requirement, a receptivity requirement, and a reactivity requirement (see §2.3):

> *Ownership requirement*: for an agent, S, to be morally responsible for a harmful consequence, UHC, the output that caused UHC must have issued from S's own processes of practical reasoning (which Fischer & Ravizza call S's "decisional mechanism");
>
> *Receptiveness requirement*: for S to be morally responsible for UHC, the output that caused UHC must have been responsive to S's patterns of reason recognition, including her moral reasons;
>
> *Reactivity requirement*: for S to be morally responsible for UHC, this receptiveness should have concretely translated into the output that caused UHC.

The second stage concerns the consequence. On Fischer & Ravizza's analysis, guidance control over the second stage of the sequence obtains when the consequence is 'appropriately sensitive' to the output. A consequence is appropriately sensitive just if: (i) it would be reasonable to expect the harm to follow from the output; and (ii) the harm occurs primarily because of the output. In including this latter requirement (ii), Fischer & Ravizza have in mind cases where one attempts to divert or minimise a threat instigated by someone else but causes a harm in the process of doing so. One of the colourful examples they give is of a person diverting a missile fired by someone else onto a smaller city than its intended target, Washington D.C. (1998, p. 116).

3.3.1.1 Low-level human agents and the control condition

Taking each part of Fischer & Ravizza's analysis, let us now consider the control that low-level human agents do and may not exercise over the system at each stage of the causal pathway to the harm. I submit that low-level human agents *do* meet the ownership requirement of guidance control. We can connect this to Aristotle's notion of control – that the source or origin of the action is the agent. Their meeting the ownership requirement is captured in the first column of Table 1/Table 1(b). The low-level agents make decisions about the deployment systems on the ground. They voluntarily decide to instigate the system's real-world agentive function; that is, they voluntarily decide to deploy the system, whether the autonomous missile or submersible, or the nurse-bot, or the ultra-high-speed financial trading system. This instigation issues from antecedent low-level human decision-making. The low-level human agents set the system in action as an upshot of their practical reasoning. There is therefore a significant respect in which *all* of the system's outputs issue from the low-level human agent's decisional mechanisms. The ownership requirement is therefore met in this respect both for UHC and FHC. Later in this sub-section, I argue that high-level human agents also meet this ownership requirement, though differently.

In many ways, arguments that humans have responsibility-grounding control of an autonomous system depend necessarily on considerations that are similar to, or can even be assimilated to, fulfilment of the ownership requirement. Himmelreich, for example, puts it thus, with respect to the low-level humans' ownership of the output: "if a were to give this order, then x would occur (in all relevantly similar situations), and if a were not to give this order, then x would not occur (in all relevantly similar situations)" (2019, p. 736). And this emphasis on the user as source of the occurrence of x, the output, is particularly compelling in the military cases, due to the inherently hierarchical nature of military command (Schulkze, 2013). Santoni de Sio and van den Hoven also maintain that the ownership condition can underscore the tracing of responsibility back to the human agents (2018, p. 8). I am in agreement with these claims but, as I shall argue in the next Chapter, tracing responsibility back to the human agents requires a deeper explanation than merely instigating the deployment of the system.

Our present concern is not the inference the 'responsibility gap' argument must make from a suspension of the Aristotelian conditions to a responsibility gap. Rather it is a consideration of the truth of the claim of the suspension of the conditions in the first place. While the ownership requirement is met, it is not clear that low-level humans meet the *receptivity* requirement of Fischer & Ravizza's notion of guidance control. What I have in mind here is that, if the facts in the world that figure in the autonomous system's model and determine its output are the *same* as those that would figure in the low-level human agent's decision-making, then there is some receptiveness of the system's behaviour to the low-level human agents' patterns of reason recognition. If the homework bot regularly and reliably picks the sources of material that Morgan would have chosen, if Myrtle takes a route that Jacinta would have chosen, there is some receptiveness here. But if low-level human agents' meet the receptivity requirement, it will be down to how involved they were, or how well-represented, in the design process.[39] Moreover, as we shall see in the discussion of high-level receptiveness to follow, given many of the computational techniques, even high-level meeting of this requirement will be weakened in many cases, so the low-level agents' receptivity will only be as good the high-level agents' receptivity upon which it depends.

In Chapter 2, I broke Row II and Row IV cases down into different ways in which unforeseen harm from the system's behaviour might come about. Row II cases comprised 'genuine accidents' and 'honest mistakes'. Row IV cases comprised 'emergent behaviour' and 'undetected error'. Clearly emergent behaviour is one pathway to UHC where the system is not receptive to low-level human agents' patterns of reasoning. To take our case of the assistive lifting robot, Tom's unusual way of speaking and loud noise overhead would not have figured as a reason in a *human* carer's thought processes about the appropriate level of force when lifting Tom from a sitting to standing position, and it certainly did not figure in Tom's reasoning. Emergent behaviour constitutes the primary case in which the receptivity requirement is suspended in general in UHC cases (and this is true also for high-level human agents, as we shall see). What of the other cases? It seems that genuine accidents, when the system's foreseen outputs cause harm largely due to other events in the environment, would also not be receptive to low-level reasoning (unless the system's mitigation strategies were in

[39] Another way in which this dimension of low-level guidance control might come about is through features such as "customisable ethics settings", whereby the low-level human agent would be able directly to choose, more or less, at least *some* of the facts in the world that would figure in the system's algorithmic processes (Thoma, 2021; Millar, 2017; Contissa, Lagioia & Sartor, 2017). But this would not extend to *all* functionality. Moreover, these proposals carry significant practical and moral concerns (Lin, 2014).

line with low-level agential intentions). Imagine, for example, an autonomous medical system that causes harm to a patient by increasing the amount of medication on the basis of her test results, but that one of those results was anomalous due to an error in sampling. An experienced clinician (low-level agent), would likely ignore the anomalous test result, but the system would not (Habli, Lawton & Porter, 2020, p. 253). A harm caused in this way, then, would be case of genuine accident in which the low-level human agent does not meet the receptivity requirement.

This brings us to honest mistakes and undetected errors. The honest mistakes and the undetected errors might be operational mistakes and errors made by low-level human agents or design mistakes and errors made by high-level human agents. Our present concern is with the former. Imagine that Jacinta deploys the self-driving car on that particular road thinking that it is within the operational design domain, but that it is in fact not. This would be an honest operational mistake on her part (Row II). Or imagine that Morgan's homework bot did in fact have an age-rating restriction function but that Morgan had not realised it was on factory settings and therefore turned off. This would be an undetected operational error on her part (Row IV). In these cases, I do not think it is correct to say that there is no receptiveness of the system's output to low-level patterns of reasoning. Rather, the system is responding, in some respect, to low-level reasoning that is wrong or misguided in some way. But if the honest mistake or undetected error were down to the high-level human agents, then the results would not be receptive to the reasons of low-level human agents.

To summarise the discussion of low-level human agents and the receptiveness requirement, this is derivative upon the extent to which high-level human agents meet the receptiveness requirement. Depending on how accidents are managed and mitigated by the system, low-level human agent receptiveness in cases of genuine accident may be substantially weakened. Receptiveness would be suspended in cases of emergent system behaviour. Harms caused by operational mistakes and undetected errors would be, in a respect, receptive to low-level agents' reasoning, but harms caused by the mistakes and errors of high-level agents would not be.[40]

[40] There are nuances that could affect these general claims, for example, if the low-level agents had been deeply involved in the design process and therefore influenced the making of those high-level mistakes.

The *reactivity* requirement falls out of the question of receptivity. It is worth recognising that the systems, unlike humans, will *always* implement the next step in the sequence unless impeded from doing so by outside interference. Even when the system is not receptive to the reasons of low-level human agents in its input-to-output processes, it will be strongly reactive to whatever reasons it is responsive to. That is, it will keep pressing on as if there is no sampling error in the test results or as if small people on scooters do not need any extra caution. Indeed, this *exacerbates* any dilution of guidance control.[41]

Let us now consider low-level guidance control over the second stage of the sequence: the causal pathway from output to UHC. On Fischer & Ravizza's analysis, guidance control over the second stage of the sequence obtains when the consequence is 'appropriately sensitive' to the output. A consequence is appropriately sensitive just if: (i) it would be reasonable to expect the harm to follow from the output; and (ii) the harm occurs primarily because of the output. To illustrate how the latter requirement (ii) might not be met, imagine, for example, in the Jacinta and Mrytle case that Jacinta does resume operational control of the vehicle and so has actual sequence, hands-on guidance control of the vehicle when the child is injured, but Jacinta is doing her best to avert the threat in front of her. Add the constraint that the cause of the child's running out into the road is her parent's negligent act. On an adaptation of Fischer & Ravizza's account of the second requirement on appropriate sensitivity, Jacinta would be excused of moral responsibility.

The former requirement (i) that it would be reasonable to expect the harm to follow from the output is an epistemic question. It concerns what one *can* anticipate. I have said I have in mind cases in which the fact of the consequence's being unforeseen *was* more or less reasonable (see §2.1). The reader might dispute this is borne out in my example cases. But we should admit *some* natural human error. No low-level or high-level human agent is epistemically perfect. It would be unreasonable to expect them to be so. So, by the lights of my own framing in Chapter 2 of the UHC cases we are concerned with, we can also say that low-level guidance control would be suspended over the second stage of the sequence.

The low-level human agent might still possess *regulative* control over the second stage of the sequence. The low-level human agent's operational situatedness means that she should be

[41] This analysis of the *reactivity* requirement also maps onto the high-level case. I therefore leave it out of the discussion of high-level guidance control below.

able to apprehend some unexpected hazards and physically intervene to prevent them. Indeed, this is a plausible duty on the user (Hevelke & Nida-Rümelin, 2015 , p. 624) and one that has been reflected in regulatory frameworks on the safe deployment of autonomous systems (Law Commission, 2021; Law Commission, 2020). The ability to intervene in this way is regulative control. Thus, we can see that even where low-level *guidance* control is substantially weakened, this does not mean that the low-level human agent would not have *regulative* control over the deployed autonomous system.

But we should also recognise that low-level regulative control may be *strict* but not *meaningful*.[42] There is a difference between a literal capacity for the low-level human agent to intervene and take over control of the system, or simply to turn it off before it can cause harm, and a practical capacity that is actually open to the agent. This capacity for meaningful control is related to the system's increasing autonomy on the three dimensions of machine autonomy. The further the system is along the first (x-axis) and second (y-axis) dimensions of machine autonomy, the less time or capacity low-level humans will have to intervene. Moreover, effective intervention requires a good understanding of the system, which I will discuss under the knowledge condition. Systems that have high degrees of autonomy in terms of independence from explicit human instruction (the z-axis) are often commensurately difficult for low-level humans to interpret quickly or well prior to optimal human intervention, and unwarranted intervention in the system's operations could well do more harm than good (Hevelke & Nida-Rümelin, 2015, p. 624). These considerations will be discussed further under the knowledge condition. But it suffices for now, for our discussion of the suspension of the control condition, that not only will low-level *guidance control* with respect to UHC sometimes be *weakened* in some respects and suspended in others, but low-level *regulative* control may well not be *meaningful*.

3.3.1.2 High-level human agents and the control condition

We can now consider whether, and to what degree, or in what respect high-level human agents exercise control over the causing of the harm, UHC. We shall follow the same

[42] Santoni de Sio & van den Hoven (2018) put forward a theory of, and design proposal for, autonomous systems enable meaningful human guidance control. My discussion here, however, is not to do with the ideal design of systems to improve the robustness of human control (although this is an important project). Rather, my focus here is on real-world ways in which control might be lost.

procedure – considering high-level guidance control over the two stages of the reasons-responsiveness sequence, and then considering high-level regulative control.

I think it is clear that high-level human agents also meet the *ownership* requirement of guidance control over the output, although what they meet the ownership requirement *for* differs in this case to the low-level humans. Low-level decisional mechanisms are the source of the system's *being operative* in the world, but high-level decisional mechanisms are the source of *how the system is operative* in the world. This too connects to Aristotle's notion of the origin or source of the action being the agent. It is also captured in the first column of Table 1/Table 1(b). High-level agents voluntarily decide to develop the system. They make decisions about how the system's agentive function is actually to be achieved, the means by which it is achieved, and what constraints are placed upon its behaviour. The system's patterns of behaviour, or its series of outputs, are of the nature that they are in the real-world as an upshot of high-level practical reasoning. There is therefore a significant respect in which *all* of the system's outputs also issue from the high-level human agent's decisional mechanisms. The ownership requirement is therefore met in this respect both for UHC and FHC.

As before, it is the high-level agents' meeting of *receptiveness* requirement that is threatened by the increasing autonomy of systems. Recall that learning-based systems have relatively high degrees of autonomy in terms of the independence of their input-to-output processes from explicit human instruction (on the z-axis). The full range of intended function cannot be explicitly specified, and much of the system's programming is implicit. Due to inherent uncertainty about the rules the system itself has extrapolated from the data, precisely how it optimises for its goals is largely unknown, even by the experts. These processes may well turn on features in the data that would not figure in human reasoning. Systems that continue to learn and adapt in the operational domain exacerbate this trend. We can therefore make the general claim that, where autonomous systems are highly independent from direct human control on the z-axis, the receptiveness requirement of guidance control in general is weakened for all concerned. It is weakened for high-level human agents and derivatively for low-level human agents.

Let us look at the receptiveness requirement with respect to Row II and Row IV cases. Clearly emergent behaviour (Row IV) is one pathway to UHC in which the system is *not* receptive to high-level human agents' patterns of reasoning, either. The designers and

engineers did not want or intend for the assistive lifting robot to respond to noises and strange cues in its the environment in the way that it did. Its doing so was not receptive to their patterns of reasoning, but emerged from a combination of unexpected interactions. It seems that genuine accidents (Row II), when the system's foreseen outputs cause harm largely due to other unanticipated facts and events in the environment, would also not be receptive to high-level patterns reasoning – unless it implemented mitigation strategies that were receptive.[43] To take the example of the village's contaminated water supply, this possibility did not figure in the high-level agents' pattern of reasoning about how the system should fulfil its agentive function – it was not therefore receptive to them.

Even so, as in my discussion of low-level human agents above, I think it would be true to say that harms caused by honest mistakes made by high-level agents (Row II) *are* receptive to high-level agents' patterns of reasoning, in some respect, just as they were in the low-level case. Take Morgan again. If it was an honest design mistake that adequate constraints were not built into homework bot to ensure that the information it provided was appropriate, then the system's provision of inappropriate material would be an illustration of receptiveness to defective high-level practical reasoning. Likewise, undetected errors made by high-level agents (Row IV). If it was an undetected design or training error that the sensors on Jacinta's self-driving car could not cope with certain manifestations of dappled sunlight, then the system's failure to detect the child in the road in these conditions would be, at least in a respect, a response to deficient high-level reasoning.

This brings us to high-level human control over the second stage of the sequence, the pathway from output to consequence. We have seen that the guidance control at this stage requires the consequence to be appropriately sensitive to the output. The first requirement (i), that it would be reasonable to expect the harm to follow from the output, has, as before, been ruled out by my focus on the reasonably unforeseen consequence. But we can still consider the second requirement (ii) anyway, that the harm occurs primarily because of the output. Imagine the first requirement is met, but the harm occurs as a result of the system going into some damage limitation procedure to ameliorate consequences of accidents that are the fault of a third party. Through an intervening set of mechanisms, rather like the counterfactual interveners of Frankfurt cases, high-level agents *can* stop the system or put it

[43] To note, if the facts were unanticipated due to honest mistake or undetected error, I would not class this as a genuine accident. Moreover, in general, I think cases of 'genuine accident' particularly warrant case-based reasoning, and that generalities about human control in these cases may not be fully feasible here.

into a 'minimal risk state' thereby exercising indirect regulative control. But, as with low-level humans, this high-level indirect regulative control may not always be effective. To take the case of the self-driving car and the child again, the vehicles' slowing down and therefore entering a minimal risk state was not quickly enough implemented. And where no human meaningfully *can* intervene – whether low-level or high-level – to curtail or reduce the effects of one of the system's outputs, human regulative control over the pathway from output to consequence is suspended.

To conclude my discussion of the control condition, I have said that it is a central premise of many responsibility gap arguments that human agents "lose control" over the actions taken by autonomous systems, or their consequences. I have considered whether this premise is true in the case of UHC. First, *direct* control is necessarily lost. Second, the picture with *indirect* control is nuanced. We can summarise as follows. Both low-level and high-level agents meet the *ownership* requirement of guidance control in all cases. Where systems have high degrees of independence from explicit human instruction, the *receptiveness* requirement of guidance control is weakened for all concerned. In cases of harms caused by emergent behaviour, it is suspended. But if harms are caused as a consequence of an honest mistake or an undetected error, the receptiveness requirement would still be met to some degree. Moreover, while strictly both low-level and high-level should be able to exercise indirect regulative control over the causal pathway from output to harm, in practice this will not always be meaningful. Di Nucci (2020, p. 59-60) and Nyholm (2018, p. 1209) hold that, because human agents have regulative control over the systems, it does not matter to ascriptions of moral responsibility whether human agents exercise guidance control. But there are plausible foreseeable cases of unforeseen harm where they have neither form of control, at least not meaningfully, at the critical harm-causing stage. To conclude, sometimes the claim about a loss of control *is* true and not because *direct* control is lost. My argument in the next Chapter is that we should not infer from the truth of this premise a responsibility gap.

3.3.2 Knowledge condition

Another premise of responsibility gap arguments is that human agents also fail to meet the knowledge condition. This claim, right from its earliest articulation by Matthias (2004, p. 182), has often been justified by reference to the inherent unpredictability of many

autonomous systems built using learning techniques. I have restricted the immediate focus of the inquiry to the unforeseen harmful consequence, UHC (see §2.1). I have further allowed that it is reasonably unforeseen, in that its being unforeseen is not particularly epistemically vicious but within the boundaries of a non-negligent lack of foresight. *Prima facie*, all of the hypothetical examples I gave in the introduction meet these constraints. If this is true, then the knowledge condition is *de facto* not met by the candidate human loci of moral responsibility in the cases of UHC with which we are concerned.

However, an evaluation of the ways in which low-level and high-level human agents might be ignorant of the system's output or its pathway to a harmful consequence is instructive. It can help us to understand how and why harmful consequences may not be foreseen by human actors in the first place and just how common such cases could prove to be. My approach in this sub-section will be to switch order, and discuss the high-level agents first, and the low-level agents second.

Challenges to high-level human agents' fulfilment of the knowledge condition come very early in the design phase of an autonomous system. Many of the challenges are captured in what has been called the 'semantic gap' (Burton *et al.* 2019). This gap arises when the designer's true intentions for the system, her ideal design blueprint, cannot fully be captured in the explicit and concrete specification that is used to build the system. The causes of this gap are: the inherent synchronic and diachronic complexity of the operational domain, for which a complete specification is not feasible to specify; the inherent complexity of the system, which may also change diachronically through interactions within the operating environment, introducing risks not foreseeable during design; and the inherent agentive function of the system – to replace a human agent in a decision-making task – which introduces a range of new functions that "have historically relied on human interpretation, ethical judgement, and lawful behaviour" (Burton *et al.*, 2019, p. 2). In addition, as we saw in Chapter 1, many systems that have high degrees of autonomy on the z-axis, in particular systems with deep learning models, are inherently inscrutable and difficult to interpret. And what is difficult to interpret is difficult to predict. Assurances that the system will perform as expected are determined at testing and validation stage. The more severe the possible harm that could be caused by the system, the more rigorous these procedures will be. But given that the systems will be employed in real-world and not simulated domains, not every possible situation it will encounter can be anticipated in advance. And when the systems

deployed are also online systems, that continue to adapt 'in the wild', these anticipations will be even more difficult. In short, there is a limit to the robustness of the high-level agents' foresight on each systems' future outputs, and there will be many cases in which the 'responsibility gap' argument's claim of an unmet knowledge condition is also true.

Let us look now at the knowledge that low-level humans may or may not have to the system, its outputs, or its consequences. As above, many systems that have high degrees of independence from explicit human instruction (the z-axis) are commensurately difficult for low-level humans to interpret. This in fact is one of the reasons low-level regulative control may not always be meaningful, because the low-level human may not know how *best* to intervene, and unwarranted intervention in the system's operations could well do more harm than good (Hevelke & Nida-Rümelin, 2015, p. 624). High-level agents may implement visual cues or dashboards that make it easier for the low-level agent to understand and interpret the system's behaviour, and as we have seen there are a suite of 'Explainable AI' techniques that are available for this purpose. But these, too, are approximations: models and plausible explanations of the underlying logic of the model. These techniques are subject to some uncertainty. It seems that for low-level human agents, too, the knowledge condition will not be robustly met.

This brings us to negligence. The knowledge condition does not exculpate ignorance due to negligence (Smith, 1983). But what constitutes the upper bound of negligence in these cases involving highly complex autonomous systems is unclear. The epistemic difficulties mentioned above make it difficult for high-level agents, including regulators, even to make probabilistic estimates about risk of harm from a system and to determine what constitute acceptable risk thresholds above which a system should not be deemed as fit for deployment. But if we accept that the UHC cases are those in no human actor is *clearly* negligent in not foreseeing these consequences and taking pains to mitigate them, then perhaps the natural conclusion we should draw in these cases is that no one is blameworthy for those unforeseen harms. I pick up this question in the next Chapter, when I consider *the locus of moral responsibility as blameworthiness*.

To wrap up my discussion of the knowledge condition, I have essentially stipulated that neither low-level nor high-level humans meet this condition due to the restriction of focus to the unforeseen harm. But a closer look reveals how, through the 'semantic gap', the human

candidate loci for moral responsibility have insecure epistemic access to the possible effects of the system's behaviour in the operating environment.

This concludes my discussion of the how and when the Aristotelian conditions are suspended with respect to the autonomous system's causing of the harm. In this Chapter, I have attempted to give a detailed consideration of this central sceptical claim. A far more nuanced picture emerges than the bold claims of a suspension of Aristotelian conditions sometimes suggest. There is a danger that the claim could be overblown and taken to apply universally to all human relations to autonomous systems. With respect to the control condition, what applies universally is that *direct* human control is suspended. This is analytically true, since autonomous systems are defined in terms of their independence form direct human control. And it is necessarily true, since human agents are not personally performing the decision-making tasks that autonomous systems are carrying out for humans. But, in terms of retaining elements of *indirect* guidance control, my adaptation of Fischer & Ravizza's analysis of guidance control also reveals that the *ownership* requirement is met by both low-level and high-level agents in all cases. The crux of the indirect guidance control problem is that as systems increase in independence from explicit human instruction on the z-axis, their *receptiveness* to human patterns of practical reasoning will become weakened. Even so, weakening is not yet suspension. But in Row IV cases of harm caused by emergent behaviour, receptiveness could be reasonably said to be suspended for human agents. The other crux is that regulative control – the capacity for human agents to intervene during deployment – will sometimes *not* be *meaningful*. And with respect to the knowledge condition, we can say that the very fact of the existence of Row I cases, which are readily conceivable, and where both the output and the harmful consequence are foreseen, belies any claim that the suspension of the knowledge condition is also universal. But our consideration of the considerable conditions of uncertainty, inherent to the systems, to the tasks they carry out, and to the domains in which they do so, shows just how *easily* an output or its consequence *could become one* that had not been foreseen – unforeseen harms are always possible. They are not rarefied cases at all.

Thus, we can see that the suspension of the Aristotelian conditions is a legitimate problem. Though this premise in 'responsibility gap' arguments is not always true, it is sometimes true, and UHC cases are such cases (in the case of emergent behaviour, and some genuine accidents, these cases qualify on control grounds and not just epistemic grounds). In UHC

cases, one or both of the necessary conditions for moral responsibility are insufficiently met by high-level and low-level human agents, *prima facie*, to warrant an ascription of moral responsibility to them. These agents would therefore *appear* to be excused from moral responsibility. A refutation of this claim of a responsibility gap must show one or both of the following: i) that these conditions are in fact met sufficiently robustly by one or more agents, even with respect to reasonably unforeseen consequences; or ii) that exceptions to their fulfilment apply. I take the latter route.

My considerations in this Chapter lead me to the conclusion that the suspension of the Aristotelian conditions in this subset of cases is true. Such cases may become more common as systems increase in independence from human intervention, oversight, and explicit instruction. But, as I shall argue in the next Chapter, there is an exception to the necessity of the Aristotelian conditions that applies here. As such, it is not the premise of the responsibility gap arguments that I object to, but the inference drawn from this premise to a lacuna in responsibility.

Summary of Chapter:

The greatest pressure on Standard View positions comes from the problem of the 'responsibility gap', which in turn rests on the claim of a suspension of the Aristotelian conditions. This suspension of the Aristotelian conditions is by no means universal, but it does obtain in some cases. Moreover, when it does obtain, it can be described in more detail as involving one or more of the following features: substantial dilution or suspension of the receptiveness of the system to human patterns of reasoning during deployment; lack of meaningful regulative human control over the pathway from output to consequence; and inherent unpredictability and uncertainty leading to unforeseen consequences. In the next Chapter, I defend the Standard View in the face of this challenge and provide a framework for ascribing moral responsibility to human agents even in these cases.

CHAPTER 4
RESPONSE TO THE CHALLENGE: A DELEGATION FRAMEWWORK FOR THE DEFENCE OF A PLURALIST STANDARD VIEW

In the fourth Chapter, I defend a delegation framework for ascribing moral responsibility to human agents in response to the problem of the suspension of, or weakening of, the Aristotelian conditions. I thereby defend a Standard View position. I argue that, irrespective of and notwithstanding any suspension of the Aristotelian conditions, moral responsibility as attributability traces back to the principal decision-makers in the complex decision to delegate to the autonomous system. The normative asymmetry entailed by this decision provides a specific tracing principle – the principal-proxy tracing principle – which warrants a responsibility attribution. I further argue that, given the specific risk committed to by these principals in virtue of their decision to delegate, moral responsibility as blameworthiness would also be borne by those principals who, even if they had not been negligent, and even if their risk analysis included a consideration of moral values, did not delegate for a reason that morally justified the risk imposition. Thus, I answer the locus of moral responsibility question.

4.1 Tracing principles for suspensions of the Aristotelian Conditions

It will be helpful to remind ourselves where we are in the argument. I have stated what I have called the *locus of moral responsibility* question (see §2.2). This question asks who is morally responsible for unforeseen harms caused by autonomous systems, and why. I set this out as a three-term relation: S is morally responsible for UHC in virtue of P. Generally, the P, in virtue of which an agent S is morally responsible for something, is fulfilment of the four conditions of responsibility. This includes the Aristotelian conditions of control and knowledge.

As we saw in the last Chapter, there will sometimes be suspensions of the Aristotelian conditions over the causal pathway to harm. Human fulfilment of the receptiveness requirement of guidance control may become weaker as systems become more autonomous on the z-axis, and human agents will not always be able meaningfully to intervene before harm is caused. Various pathways, denoted by Row II and Row IV cases, can lead to unforeseen harm, where the knowledge condition is suspended precisely because of its being

unforeseen. A survey of the ways in which unforeseen harmful consequences might come about – accidents, honest mistakes, surprising and emergent behaviour, undetected error – shows how common such cases could be.

In this Chapter, I argue that, while the suspension of the Aristotelian conditions is sometimes true, discernible human agents are still eligible for moral appraisal. That is, we should not infer a 'responsibility gap' from the suspension of the conditions. These cases fall into a class of exceptions to the Aristotelian conditions that are covered by a tracing principle which traces responsibility attributions back to a point at which these conditions were not suspended: if one is responsible *for* a loss of control or ignorance one cannot be excused of responsibility *by* that loss of control or ignorance.[44] The stock example is the drunk driver. If a driver hits a child when driving drunk – and therefore not controlling her action nor really aware of her action – the hitting of the child is still attributable to her in a way that is necessary to serve as a basis of moral appraisal. The child's injury is attributable the driver because of her antecedent decision to do something that would incur the suspension of the Aristotelian conditions at a later time. She chose to relinquish future control and awareness. She is not excused of responsibility for the consequences of that choice.

In fact, 'being responsible for the suspension of the Aristotelian conditions' can be understood in two ways: actually *bringing about* the loss of control or ignorance; or deciding to act *notwithstanding* the fact the conditions are or will be suspended. Our drunk driver does both: the former by getting drunk; and the latter by getting in the car and driving. Both senses apply also to low-level and high-level human agents with respect to the system's output and its pathway to the unforeseen harm. The suspension is circumstantial upon decisions about its degree of machine autonomy on the three dimensions I described in Chapter 1; those decision-makers therefore *bring about* the loss of control and epistemic access. And in the cases we are concerned with, the antecedent human decisions were made *notwithstanding* the fact that the conditions would be. The classic catch, therefore, applies to human agents here, too.

But an analysis of the underlying structure of the relation between human agents and autonomous systems yields a particular form of this tracing principle for ascriptions of moral

[44] Tracing principles are widely discussed in the moral responsibility literature. See: Nelkin & Rickless 2017; Shabo 2015; Khoury 2013; Khoury 2012; Fischer & Tognazzini 2009; Vargas 2005.

responsibility for harms caused by autonomous systems. The tracing principle I discuss in the next three sections does not just concern diachronic responsibility (Khoury, 2013), such as the drunk driver case, whereby one's moral responsibility for a consequence at t2 traces back to one's moral responsibility for an antecedent decision at t1. It also concerns responsibility that extends through other agents (or autonomous systems), whereby one's moral responsibility for a consequence at t2 traces back to an antecedent decision at t1 to *delegate to another agent* (or system), whose subsequent action (or output) causes that consequence at t2.

This furnishes specific properties P that override any suspension of the Aristotelian conditions with respect to the immediate causal antecedent of the harm, and in virtue of which attribution of the harm to the delegating agents would be warranted. The principal-proxy tracing relation, as I call it, and which I discuss in §4.2 below, has normatively asymmetrical features which provide a moral basis or justification for attributions of moral responsibility to the principals, notwithstanding any loss of control or knowledge during the deployment of the system. Moreover, the particular risk to which the principals commit when deciding to delegate yields a morally demanding or strict threshold for responsibility as blameworthiness, which I discuss in §4.3 below.

The grounds for ascribing moral responsibility for harms caused directly by autonomous systems arise from viewing the notion of delegation as central. My approach has its roots in the work of Di Nucci (2020) and Di Nucci & Santoni de Sio (2014). My conceptualisation of the delegation framework is independent, but the underlying commitment is the same: that understanding the relation between humans and autonomous systems as a relation of delegation is descriptively accurate, and explanatorily and normatively powerful. Most responses to the problem of responsibility incurred by autonomous systems are grounded in some kind of relation between agents (Floridi 2016) or on a combination of relations and roles (Coecklebergh, 2020a; Nyholm, 2018; Schulkze, 2013). An explicit delegation framework does, too, but it goes further in providing a principle to explain the weighting of responsibility with the principal. I show that this enables us to identify fairly the *loci of moral responsibility as attributability* in even the most intractable cases of genuinely unforeseen harm. This shows the delegation framework has greater capacity for ascribing moral responsibility than a standard instrumentalist theory.

4.2 The principal-proxy tracing principle

The instrumentalist theory of technology holds that all technological devices are tools, and that computational systems, including autonomous systems, "like all other technological artifacts, [will] always be instruments of human value, decision-making, and action" (Gunkel, 2017, p. 3). But what is it about this position that grounds an ascription of moral responsibility? One obvious answer is the respective roles and duties of toolmakers and tool users (Boden *et al.*, 2017). But if autonomous systems are tools, which I think is hard to dispute, they are tools with a specific property, which distinguishes them from 'mere' tools and standard rule-based automata. The latter classes of tool are *used* by the human, the former kind *replace* the human (Di Nucci, 2020, p. 69). Now, it might be thought that bike locks replace the human hand holding onto the bike, and that washing machines replace the human launderer, and that tractors replace teams of human farm labourers. So let me be more specific. Autonomous systems replace the human in cognitive tasks - tasks that would require decision-making in the human. All the decision-making in the former class of tools is done by a human. To replace the human decision-maker is the purpose – the agentive function – of an autonomous system. We might say, inelegantly, that the agentive function of an autonomous system is to have decision-making function transferred to it. But this can be expressed, without loss of sense, as a handover of decision-making function. That is not to say that the system itself makes decisions in some full-blooded human-like sense. It is rather to say that it carries out a task that would require decision-making, including interpretation and judgement, in a human fulfilling the same role.

My first claim is that delegation is an accurate description of what happens when someone hands over decision-making function to another agent, which or who replaces the person in that task. As Di Nucci puts it, "when you delegate, the recipients act *on your behalf* (or *in our place*). This is the difference between cooking – using all the tools, instruments, and machines that this requires – and having someone cook for you" (2020, p. 69). Moreover, you delegate for reasons: to get better results, to spare resources, to save precious time. A reader sceptical of my claim of descriptive accuracy might point to agential differences between human proxies and artificial proxies. The objection might be that autonomous systems simply do not have the requisite agential capacity to be delegates. And perhaps my insistence on the words 'output' instead of 'action', and 'system' instead of 'agent', and 'decision-making function' instead of 'decision-making', play into the hand of this objection. Human delegates

consent to their mandates, and act intentionally when they carry them out. Autonomous systems do not; so perhaps we do not delegate to them at all. Perhaps, rather, the deployment of autonomous systems is an extension of automation, and we will change and adapt our behaviour to take advantage of the ease and efficiency that these new extensions of automation provide.[45]

A difference in kind between the agential capacities of an artificial and a human proxy is certainly relevant to the question of whether moral responsibility is always *exclusively* borne by human principals (that is, whether our Standard View position should be hard or soft). I discuss this in the next Chapter. But the objection does not impugn the legitimacy of the claim that handovers of decision-making function to the system are nonetheless acts of delegation. There is a salient difference between automated rule-based systems and autonomous learning-based systems which can be used to respond to the extension of automation objection. With automated systems, human instruction *pre-determines* the operation of the system. This does not constitute independence from human rational control. By contrast, with autonomous systems, human instruction *under-determines* the actual operation of the system. Here, the independence starts to come in, and with it the third dimension – what I have called the z-axis – of machine autonomy. This under-determination is not an extension of pre-determination; it is a meaningfully different thing; it is a paradigm shift.

If we were analysing the systems in terms of the x-axis and the y-axis alone, then there is no salient difference between automated and autonomous systems. After all, even my washing machine operates without much interference from me (x-axis), and I certainly don't watch over it much (y-axis). But humans have robust rational control of the washing machine's behaviour: its intended functionality is explicitly specified, and while it may have a fault, it will not start to implement unforeseen outputs, at least not unless there is some truly freak event. The high-level human agent antecedently exercises this rational control, during the design and development of the system, and the low-level human agent exercises it in real time, when making use or deployment decisions (what water temperature to select, what spin cycle, and so on). By contrast, autonomous systems have increasing independence from explicit human instruction (which is the z-axis). These selections are far more *implicitly* and *incompletely* guided by the high-level agent, and the results may even be surprising to the high-

[45] I thank Vincent Müller for raising this objection with me.

level agent. Here, the high-level agent does not fulfil the decision-making function *ex ante*, but rather creates a system that will be able to fulfil the decision-making function *itself*. I submit that this 'being able to fulfil the decision-making function itself' is what marks a difference, rather than simply an extension, to normal automation. As such, autonomous systems are not an 'extension of automation' as the objection maintains. To repeat, my purpose here is not to say that this capacity to fulfil decision-making function itself confers any particularly rich properties on the autonomous system, such as voluntary or intentional agency. There is no anomaly in saying that it could fulfil this function entirely mindlessly. I think a pressing responsibility question arises even with systems of minimal agency. And I do not think it would be untrue to say that we delegate to the system even if it is *not* really making decisions in some full-blooded way. What matters is that it fulfils the *function* of doing so, and specifically that *it* fulfils this function of doing so - during its operation. We delegate to *it*.

Second, the framing of the relation between humans and autonomous systems as a relation of delegation is also explanatorily powerful. It provides a clear explanation for the loss of direct control and for the weakening of the fulfilment of the receptiveness requirement of (indirect) guidance control. Indeed, suspensions of the Aristotelian conditions are integral to the delegation relation. When Jacinta gets in her self-driving vehicle and tells it where she wants to go, from that point on she hands over subsequent decision-making function to the vehicle. And it is precisely because she does so that she no longer has guidance control over the driving task, and the myriad outputs that comprise it. When my boss delegates to me the task of arranging the meeting, she no longer has guidance control over – she is no longer controlling – my process of doing so. She can no longer enforce or be sure that my actual (albeit quite pedestrian) choices will be receptive to her reasons in the moment that I am making those choices, because I am doing it and not she. Of course, she may give me her preferences in advance, strongly encourage me to be receptive to her reasons, or reprimand me if, for example, I choose a room with no natural sunlight or fail to ensure that there are vegetarian options for the meeting participants. My boss may step in and take over, but insofar as, and for the duration that, she transfers decision-making function to me, she is no longer controlling my decision-making. This maps onto what happens when we delegate to an autonomous system. As I will discuss below, some delegates are 'free' and others are 'bound' according to the discretion and initiative accorded to them, and delegates at the upper ends of the continuum may well be the more powerful party in the relationship (Di

Nucci 2020, p. 68; Feinberg, 1970, p. 226). Autonomous systems, I argue below, are bound delegates or proxies. But the delegation relation has additional explanatory power because the relation remains structurally the same across a wide range of properties that an autonomous system could instantiate.

The third claim I make about the delegation relation is that it is normatively powerful. The toolmaker/tool user framework struggles to deal with precisely the cases with which we are concerned: where the fact that the harm was unforeseen was not necessarily negligent, and where there is a suspension of the Aristotelian conditions. To ascribe moral responsibility to toolmakers and tool users in these cases would be on the basis that it is part of their role to absorb this responsibility not because they necessarily deserve it. But the delegation framework, and the principal-proxy tracing principle, provides a moral justification for the ascription of moral responsibility in these cases: it explains why such an ascription would be deserved. In the next two sub-sections, I explain how the delegation framework allows us to determine the *locus of moral responsibility as attributability* and the *locus of moral responsibility as blameworthiness*.

But first, I think we should look at where delegation begins. It follows that, given that the decision to deploy just is to realize the system's agentive function, and its agentive function just is to be delegated to, the decision to deploy an autonomous system just is the decision to delegate to an autonomous system. This decision to delegate is complex, to say the least; it is a decision of "exponential, multi-level complexity" to paraphrase Di Nucci (2020, p. 52). Below, I impose some order on this complexity and offer a description of which principals make relevant decisions at which decision-points within this multi-levelled decision to delegate. I should add that, in the description that follows, I reduce some of the complexity by referring to each kind of human agent as an individual, even though at every level of the decision to deploy, but particularly at the second and third levels, the low-level principal or high-level principal will invariably be individuals working in teams. To illustrate my points, I draw on the characters of the hypothetical examples in the Introduction.

I divide the multi-levelled complex decision to delegate into three levels.

The decision to delegate at the **first level**, closest to the actual output of the system, is made by the low-level human agent. This is Jacinta, or the military commander, or Morgan. At this

level we have the immediate instigation of the delegated task, the point of handover to the autonomous system. The agent of this decision point I call the 'low-level principal.'[46] It will always be the case that *someone* (or some discernible set of people) decides actually to initiate the transfer of decision-making function for some particular task to the system in the operational domain.

This decision is not made in isolation, however. It will be informed by assurances and authorisation given by high-level agents – the regulators – that the system is fit to replace humans in the execution of decisional tasks in the intended domain, according to an acceptable risk threshold. This is the **second level** of the decision to deploy. It is made by what I call the high-level principals, removed from immediate delegation. These agents were not mentioned in the hypothetical examples, apart from the tacit reference to the vehicle's approved operational design domain (ODD) in the first case. But in many societies, such agents will be regulatory agencies and public bodies, which set or ensure compliance with standards, regulations, and risk thresholds. Some domains, such as health care or automotive, are highly regulated; others, such as agriculture, less so.

This connects to the **third level** of the decision to delegate. The standards and regulations set parameters on the activities of developers and are action-guiding in the creation of the autonomous systems. At the third level decision point, another group of high-level principals 'bind' the autonomous system to fulfil its agentive function as a proxy for the low-level principal in the operational domain. Here we have the engineers of the self-driving car, the ML-enabled missile, the personal robotic homework assistant, the assistive lifting robot for frail users. I have mentioned that delegation obtains on a continuum of initiative and freedom. I think that we should understand the developers' activities, of setting the parameters of the model, and setting its goal and meta-algorithms, and curating its data, and testing and validating it, as 'binding' the autonomous system to fulfil its agentive function as a proxy for a human user in the operational domain. The low-level principal who makes the decision to deploy an autonomous system at the first level makes the decision to deploy an operational proxy that has been bound (at the third level) and approved (at the second level) by high-level principals.

[46] Note that the sets of low-level and high-level human agents, respectively, have more members than the sets of low-level and high-level human principals.

The autonomous system replaces the low-level principal in the world, but the binding by the high-level principals means that the indirect guidance control of the system that *does* obtain traces to the third level of the decision to delegate. We can now make the following generalisation. Operationally, the autonomous system is the proxy of the low-level principal (or, in other words, the system is the low-level principal's operational proxy). Rationally, the autonomous system is (primarily) the proxy of the high-level principal (or, in other words, the system is the high-level principal's rational proxy). More specifically, it is the rational proxy of the developer as constrained by the regulator. Not only is there a delegation relation, but autonomous systems are proxies to two principals.

There is, therefore, a deep interconnectedness between the low-level and the high-level principals in the multi-layered decision to delegate to the system. It might seem, in fact, that this interconnectedness points to irreducibly shared agency in the decision to delegate. This is, I think, true. But it is not to say that discernible principals cannot be ascribed moral responsibility for their part in the complex decision to delegate to an autonomous system. To recall, in my discussion of human guidance control in the previous Chapter (see §3.3.1), I claimed that the ownership requirement of guidance control was met by both low-level and high-level agents, and specifically, we can now say, by low-level and high-level principals. But it is clearly met differently in each case. The source of the self-driving car being out on the road that sunny day with minimal supervision was the low-level principal, Jacinta, but the source of how it behaved in the operational domain was the high-level principals. At the point of failing to detect a child in the road, the system was, we might at least allow, receptive neither to Jacinta's nor to the high-level principals' reasons at the time, and neither low-level nor high-level principal could meaningfully or effectively intervene and stop the harm. The central point of this Chapter is that the delegation relation helps us to see how such suspensions of control do not inhibit morally justified ascriptions of moral responsibility either to Jacinta or to the high-level principals.

Let me here state *the principal-proxy tracing principle* that the delegation relation yields. I defend this principle in the next section, when I discuss what I call the 'normatively asymmetrical internal criteria' of the relation:

> *Where agents S are decision-makers at one of the three levels of the decision to delegate to an autonomous system, i.e. are principals, the consequences of the proxy's outputs are as morally attributable to S as if S had acted personally, even though S did not literally act personally in the immediate production of the harm. By 'morally attributable', I mean S is fairly open to moral appraisal for these consequences.*

This constitutes an exception to the necessity of the fulfilment of the Aristotelian conditions at all stages of the sequence leading to any harms caused by the autonomous system.

4.3 *Loci of moral responsibility* as attributability

Given its ubiquity in our daily lives and worldly affairs, it is curious that the relation of delegation and its implications for responsibility are not more discussed in the wider philosophical literature (Di Nucci, 2020, p. 71).[47] Contemporary philosophers of action are far more exercised by the question of action *with* another agent than by the question of action *for* another agent. But there are some distinguished exceptions to this rule, where the phenomenon of an agent acting for another agent is discussed variously under the names of secondary agency (Copp, 1980; Copp, 1979), vicarious agency (McMahan, 2010), substitute agency (Kamm, 2008), and proxy agency (Ludwig, 2017; Ludwig, 2017). Each have a different nuance. Secondary action, for example, is particularly concerned with a group acting through an individual. Some of these accounts pay an explicit debt to the Hobbesian concept of an "artificial person" in the *Leviathan* (Part I, Ch. XVI) – one who acts in the name of another, or stands in the place of someone else, as distinct from a "natural person" who acts for herself. These maintain that the idea of an artificial person can be extended to relations in professional, business, social, and military affairs (Wolgast, 1992; Copp, 1980).

The reason the delegation relation weights responsibility with the principal is because it a normatively asymmetrical relation. Drawing on the various discussions in the literature, I take there to be the three internal criteria of the delegation relation. By 'internal criteria', I

[47] Expressed concretely in terms of 'delegation' and in explicit connection with technology, the notion of delegation is largely exclusive to the work of socio-technical scholars such as Latour (2005, 1994, 1992). Interestingly, those arguing in this vein, such as Waelbers (2009), tend to describe delegation more as creating the responsibility problem than as providing a framework for its solution. But expressed less concretely – more in terms of proxy agency, secondary agency, and vicarious agency, and without any connection to technology – analytic philosophy can be mined for the notion of one agent acting for or in the place of another.

mean the essential internal properties of the relation between principals and their proxies; we might call them the constitutive properties or features of the delegation relation.

The first property is the initiation or *instigation* of the event by the principal (Di Nucci, 2020; Kamm, 2008). Clearly, this occurs at the first level of the decision to delegate, which is made by the low-level principal. I understand instigation to be non-accidental initiation of an event; the principals *mean* to hand over decision-making function to an autonomous system. Instigation is asymmetrical because the direction of fit is *from* the principal *to* the proxy. Role responsibility for carrying out the task, for replacing the human, is transferred to the autonomous systems. But reasoning at this decision-point is not made by low-level principals in a vacuum. It will be informed by assurances that have been given of its expected behaviour, which should include indications of its risk thresholds in intended contexts, and the constraints that have been built into it. The high-level principals play a supportive and enabling role in the instigation of the transfer of decision-making function on the ground.

It might seem that instigation alone can do all the work of grounding responsibility as attributability. The principals are, by virtue of the instigation, authors of the fulfilment of the task. This has already been established, in fact, in my acknowledgement that the 'ownership' of the output and its effects is possessed by low-level and high-level agents. But a focus on instigation alone as the source of moral responsibility can become arbitrary. I am surely not responsible for *all* the consequences of things I have instigated. I may instigate a cricket match, but that does not mean that moral responsibility should be attributed to me for the fielder's being knocked unconscious by the ball. If things do not pan out as expected, it is not clear that instigation alone can explain why I am on the hook for it. The next two features – substitution and authorisation – provide this explanation. They entail that the principals are responsible for the harm *as if* it were done personally by them; it is done in their name (Feinberg, 1970, p. 227). This is what one signs up to when one decides to delegate to a proxy. In so doing, one commits to the normative consequences.

So, the second property of the relation is the proxy's *substitution* of the principal in the execution of the task (Di Nucci, 2020; Waelbers, 2009; Kamm, 2008). According to the generalisation I made in the previous sub-section, *operationally*, the autonomous system

substitutes the low-level principal.[48] That is its agentive function. But it is the high-level principals who determine, even if only incompletely and implicitly, how it does so – *how* it fulfils its decision-making function. Though the system's input-to-output process in the real-world cannot be fully specified, as we have seen, the result is still a representation of the decision-making of developers. Inevitably, then, there is a sense that rationally, the autonomous system if not substitutes, certainly represents, the high-level principal. This stretching of the substitution feature underscores my choice of the word 'proxy' for the delegate. The system is a proxy with two principals. The reader might wonder why I have not mentioned – why I did not mention, at the third level of the decision to delegate, the clear fact that developers not only bind but create the system. Creation is important, but it is the constraining of the system such that its substitution of a human is determined to be safe and appropriate that is most salient to the decision to delegate.

When the substitution feature is understood in connection with the third feature of a delegation relation – authorisation – the moral justification for attribution of responsibility with human principals becomes clear. The proxy is *authorised* by the principal (Wolgast, 1992; Copp 1980; Copp, 1979, Feinberg, 1970; Pitkin, 1967). Recognition of this property, in particular, has deep Hobbesian roots, whose artificial persons acted *in the name of* another. This is what grounds the sense that principals act through their proxies, and that, even if they do not physically cause the harm, the normative consequences are directly attributable to them. The explicit points of authorisation are at the second and third levels of the decision: the high-level principals who approve the system; and the high-level principals who confirm it has been satisfactorily 'bound'. But depending on context, authorisation may be explicit at the first level, too (for example, in a military context); and if not, it will be tacit at the first level, evident in the very fact of witting handover to the system (for example, the personal user of a self-driving car). It is becoming clearer that the decision to delegate is indeed a complex and interconnected one, since each of these features of the delegation relation are more strongly or explicitly instantiated by the low-level or the high-level principals, respectively, depending on the decision-point in question.

[48] This is not quite a hard and fast rule. In the case of the autonomous missile and the assistive care home robot, the system takes the place of a different human who the principal would otherwise have delegated to. So sometimes the artificial proxy substitutes the human principal, but where the principal would not have undertaken the task herself, the proxy relaces a human operator, who would have undertaken it.

My argument is that the normative asymmetry of the delegation relation, as contained in these three properties or features – instigation and authorised substitution - is sufficient fairly to ground the attributional responsibility for the consequences with human principals. Thus, the delegation relation yields the principal-proxy tracing principle, which enables us to answer the question of the *locus of moral responsibility as attributability*.

More strongly, the principal-proxy tracing principle traces the responsibility attribution back to the principals *irrespective* of whether the harm was foreseen or unforeseen, and notwithstanding a suspension of the Aristotelian conditions during the immediate causal production of the harm.[49] The autonomous systems do not carry out the task in the same way the principals would. The input-to-output processes that are the causal antecedent of the harm are not necessarily responsive to the principals' reasons. And principals may not have meaningful regulative control, which will often be due to epistemic constraints on understanding the system and anticipating its environment. Given that these suspensions do obtain, the tracing principle therefore constitutes an exception to the necessity of the Aristotelian conditions. Indeed, this suspension is to be expected with respect to the execution of the delegated task, because it is inherent in the substitution property of the delegation relation, and the fact that the principals are not literally performing the task personally.

The outcome of my discussion here can be expressed in terms of the three-term relation:

> *The delegation relation (specifically, its three normatively asymmetrical internal features), which yields the principal-proxy tracing principle, provides the property, P, of the relation the principals S bear to all consequences of deployed autonomous systems' outputs, in virtue of which the attribution of moral responsibility to S for UHC is warranted.*

By tracing responsibility as attributability back to more than one class of principal, I am defending a notion of shared human responsibility for these harms. This makes my defence what I have called a pluralist Standard View position: more than one kind of human principal is morally responsible.

[49] This now helps us to answer the benefit-of-the-doubt Row II and Row IV cases (see §2.1) – these principals would *still* be open to moral appraisal.

4.4 *Loci of moral responsibility* as blameworthiness

Whereas attributional responsibility concerns the principals' relations to the autonomous system, and from thence for the consequences of the system's behaviour, blameworthiness concerns the principals' relations to the people harmed (see §2.3). Our purpose now is to determine whether, and when, the loci of moral responsibility *as attributability* – the principals to whom the harm is morally attributable – would also be the loci of moral responsibility *as blameworthiness*. According to my stated in procedure (see again §2.3), it is only the determination of both of these questions that will provide a complete answer to the *locus of moral responsibility* question.

Central to determinations of responsibility as blameworthiness is the fact that the decision to deploy an autonomous system, which is the decision to delegate to an autonomous system, is also the decision to commit to the subsequent risks, which include the exposure of others to risk or the imposition of risk upon others. This strikes at the heart of the principals' relation to the people harmed.

As Möller, Hansson & Peterson note, 'risk' in itself is not a very clear concept, used variously to mean an unwanted event which may or may not occur, the cause of such an event, the probability of such an event combined with its severity, or the conditions under which a decision is made (2006, p. 421). All of these senses apply to the risk of harm from an autonomous system. By 'commit to risk', I mean, for the most part, the more general commitment to the possibility of harm to people, to moral patients, from the outputs of autonomous systems. Depending on the tasks for which the systems are deployed, the harm may be physical, as it was in the case of the small child hit by the vehicle, or in the case the elderly care home resident, Tom, wrenched by the assistive lifting robot. But the harm may also be psychological, as it was with Morgan, or an injustice, as it was for villagers at the foothills of the mountain. Cases in which autonomous systems have discriminatory outcomes – due to the inherent vulnerability of ML-models to bias, and the historical discrimination that may be encoded implicitly in datasets – are also harms, and may compound other harms: imagine, for example, that the reason the vehicle's front sensors could not detect the child was because they did not discern her dark brown skin from the shadows cast by the trees on the road, where a light-skinned child would have been detected more quickly, giving the system more time to stop. This would be a double harm to the child:

an injury, and an injustice. Not all harms are susceptible to the same risk assessment and same risk thresholds. The risk of moral harms requires that more values, such as questions of equity and whether the risk-exposed individuals benefit from the deployment, are included in assessments of morally acceptable levels of risk (Hansson, 2003, p. 302).

We have seen that the development and deployment of autonomous systems takes place in conditions of considerable uncertainty. This underlies the suspension of the knowledge condition, as we saw in §3.3.2. It is often this uncertainty that drives the pathways in both Row II and Row IV cases towards unforeseen harms, and makes it unclear at what point ignorance of the consequences is culpable. As we have also seen (§2.1), decision-making under uncertainty is different to decision-making under risk. With decision-making under uncertainty, not all outcomes are determinable beforehand. With decision-making under risk, outcomes can be assigned reasonably strict probabilities. In practice, decisions made under uncertainty are treated as if they are made under risk in order to set actionable risk thresholds: "decision-makers must treat risks categorically, as being above or below an action threshold" (Fischhoff, Kadvany & Kadvany, 2011, p. 37). Within the multi-levelled decision to delegate (which is also a commitment to the risk of so doing), it is the high-level principals at the second level – the regulators – who play the dominant role in determinations of morally acceptable risk. Even so, these questions inform and are informed by 'binding' of the systems at the third level, and they inform immediate handover to the system at the first level. As before, there is an intricacy of relations between the three kinds of principal to which responsibility as attributability traces. But at each level, given that the purpose of the decision-making at each salient decision point is to delegate to the system, the commitment to risk that this entails is implicit in the decision-making at each point. When, for example, the high-level principal at the third level decides that the vehicle's sensors do not require any more validation, and that the system is appropriately 'bound' in this respect, this is a commitment to the risk of the potentially harmful consequences of that decision, because of the importance it plays in the overall functioning of the artificial proxy to which decision-making function will be handed over at the first level.

In the second and third Chapters, we saw that there are plausibly foreseeable cases that do not occur within the accepted risk thresholds because they instantiate a feature that undermines the assumptions of the accepted risk threshold, or show it to be incomplete or inaccurate. These are Row II and Row IV cases in which unforeseen harmful consequences

arise. I argue that the delegation framework also extends to these cases, and therefore also to cases of harm caused by emergent behaviour of the system, as to well as mistakes and errors that are reasonable on the part of normal, conscientious decision-makers. Through the delegation relation, which yields the principal-proxy tracing relation, responsibility as attributability traces to the principals irrespective of whether the harm was foreseen. In all cases, they are open to moral appraisal for it through their role in binding, approving, and handing over decision-making function to the autonomous system. The fact that the harmful consequence was unforeseen makes no difference to responsibility as attributability.

The question I now pose is whether, by the lights of the delegation framework, these attributionally responsible agents are also blameworthy for the unforeseen consequences, or for committing to the risk of delegation notwithstanding the possibility of such consequences. It may be that, even if the thing for which moral responsibility is sought is a harm, and it is morally attributable to an agent, she is still not blameworthy for it – if the decisions that form the basis of her moral appraisal were demonstrably ones for which she had a strong moral justification. As Watson puts it: "in general an excuse shows that *one* was not to blame, whereas a justification shows that one was not to *blame*" (2004, p. 224).

We can decompose the risk of harm committed to by the decision to delegate as follows. The commitment to risk of harm is three-fold. It is of the risk of the *harm itself*. It is of the risk of the *harm imposed in that way*. And it is of the risk of the *harm imposed for that reason*. Below, I discuss each of these in turn.

First, there is the risk of *the harm itself*. As we have seen, autonomous systems carry out critical tasks in the living world. Some risk of harm to life, liberty, and wellbeing is an ever-present possibility, even if the particulars are unforeseen. In some cases, acceptance of the risk of the harm will be straightforwardly blameworthy because it falls below a reasonable standard of care towards those upon whom the risk is imposed. Falling below a reasonable standard of care, or failing to conform to a required standard, is what I take 'negligence' to mean (Turner, 2019, p. 84).

There are two ways in which this might be so. The first is that the principals have consciously fallen below this standard and are aware of the subsequent risk of harm. Imagine, for example, that manufacturers and engineers deliberately remove a critical safety mechanism

to achieve better system performance and know that doing so will increase the risk of harm. Or imagine that a user consciously misuses a system in a hazardous way, in awareness of and despite the risks. Or take a case in which all of the principals know that the agentive function of a particular system is to meet a morally impermissible goal (e.g. torture), or that they are aware that, while the goal is morally permissible, meeting it demands a procedural justice that could only be met by a fellow human agent. Such cases would be obviously blameworthy. They would also be out of the initial scope of my inquiry since the harms incurred would not be unforeseen (see §2.1).

The second way in which principals may fall below a standard of care is by failing to realise that their conduct is unreasonably risky. According to some scholars, a hallmark or necessary condition of negligence is the lack of a conscious element (King, 2009, p. 578; Zimmerman 1986, p. 200). On this reading, if this inadvertent imposition of risk is one that a reasonable person would have recognised and refrained from, it amounts to a culpable breach of a duty of care. Impositions of this kind of harm would also be straightforwardly blameworthy. Indeed, these would also be out of the initial scope of my inquiry since the harms incurred would not be non-negligently unforeseen (again, see §2.1).

A commonly held view in the philosophical and legal debate that an agent must *have* a duty of care towards another agent in order for it to be possible to *breach* that duty and hence act negligently (Raz, 2010, p. 6). Since my concern is with the locus of moral responsibility, we can refer to the duty of care as a moral duty. Plausibly, not all risk-imposers involved in the complex, multi-levelled decision to delegate to an autonomous system in fact do have a moral duty towards the risk-bearers or risk-exposed. It is not clear, for example, that individual designers far upstream in the development process have a duty of care towards risk-bearers who are very distant from and possibly unknown to them. It seems true, for example, that not all individuals involved in the design of such a system – for example, the research computer scientist whose breakthrough discovery was implemented in the system – bear a moral duty to all people harmed by such systems. And there are important questions about the temporal scope of such duties even where they do more plausibly apply – for example, does the duty extend to future generations and, if so, how many? But it is worth emphasising that the harm for which we seek responsibility is one directly caused by the autonomous system: it is not then about harm to future generations, or even those simply indirectly affected. Moreover, the delegation relation is what underscores attributional responsibility

and, though many risk-imposing agents will be at a distance from the harmed individuals, the harmed individuals will not be at a distance from the system itself – and the system itself is a proxy for its principals. This raises three points. First, the importance of emphasising that the risk-imposing agents we are concerned with are not high-level and low-level *agents* (the research computer scientist could be construed as a high-level agent), but rather a subset of these agents: the high-level and low-level *principals* (the research computer scientist is not among those 'binding' the system for the purpose of delegation). Second, I think it is true to say that being a principal itself confers a moral duty towards others to meet a standard of care. This can be made clear by remembering, third, that, by the normative asymmetry of the delegation relation, the direct consequence of the proxy agent's behaviour is as attributable to the principal as if the principal had caused that consequence personally. As such, the possibility of negligence applies to many more cases than one may initially expect: it applies to *all* principals. Even so, the central focus of our *locus of moral responsibility* question – and our particular focus for this question of responsibility as blameworthiness – is responsibility for harms that do not trace to negligence.

In many foreseeable and plausible cases, there will not be obvious blameworthiness. In the hypothetical examples given in the Introduction, no one was discernibly at fault, although there were some intuitions that human fault did obtain in some cases: Jacinta was rather lackadaisical; and perhaps the sensors had not been rigorously enough checked to warrant approval within the context of the whole system; and the military intelligence could have been better, for example. But assuming that these flaws are within the scope of forgivable human error, all things considered, what are we to say about the *locus of moral responsibility as blameworthiness* in such cases?

This brings us to the second aspect of the commitment to risk: the risk is of a *harm imposed in that way*, through delegation to an autonomous system. First, it is the imposition of a risk of harm from a machine, which in itself may be morally and psychologically damaging to the harmed individual. Second, it is a commitment to risk from a system which, due to the implicit programming that under-determines its behaviour in the operational domain, has increased scope to act in unforeseen ways. Third, as I outlined when I discussed the explanatory power of the delegation relation, it is a commitment to the risk of a harm from a system over which direct control and direct epistemic access will necessarily be lacking. In addition, as we saw in Chapter 3, there may well be severely weakened indirect guidance

control, and possibly not even meaningful regulative control over the pathway from the output to the harm. Fourth, because the system is not carrying out the task as a human would, by the decision to delegate, the principal instigates a whole new range of possible, undesirable consequences – done in one's name but which may not have occurred if one had not engaged an artificial proxy.

These features of a risk of harm imposed in that way make moral considerations about the acceptability of the risk particularly salient. These include the values of things such as the equity of risk distribution, such that the cost does not fall disproportionately on the person exposed to risk, and the power of the risk-bearing agents to resist the risk-imposition (Hansson, 2018). These values matter as much as the probabilities given to the possible outcomes of an autonomous system's behaviour in the operational domain. For the commitment to risk not to be blameworthy, therefore, I posit that, in addition to the requirement of not falling below a moral standard of care, these moral values such as fairness need to be included in the assessments of risk that are committed to at the three levels of the decision to delegate to the autonomous system. To be clear, however, these ethical risk assessments should *supplement* the standard feature of weighing severity and likelihood of risk, and of rendering some risks too severe for the risk to be tolerable. This is relevant to the third aspect of the commitment to risk, for it is a constraint that stops extremely severe risks of harm and indeed the imposition of severe harms themselves being deemed permissible considered so long as they are intended ultimately to serve morally desirable goals.

As Hansson notes, "risks are inherently connected with interpersonal relationships. They do not just "exist" as free-floating entities; they are taken, run, or imposed" (2003, p. 302). I argue that, in order for the decision to deploy an autonomous system (and hence the risk-imposition committed to when doing so) to be non-culpable, it must, in addition to the two previous requirements, also have a morally good justification for being imposed in the first place.[50] This is the third aspect of the commitment to risk: it must be imposed for the right reasons, such that the purpose it serves justifies the risk committed to. It is this aspect of the commitment to risk - to a risk of *harm imposed for that reason* – that helps us to determine whether the decision to delegate is or was morally justified.

[50] To be clear, however, these ethical risk assessments should *supplement* the standard feature of weighing severity and likelihood of risk, and of rendering some risks too severe for the risk to be tolerable. This is relevant to the third aspect of the commitment to risk, for it is a constraint that stops extremely severe risks of harm and indeed the imposition of severe harms themselves being deemed permissible considered so long as they are intended ultimately to serve morally desirable goals.

A morally neutral case – where the goal the delegation serves is not in and of itself impermissible, such as driving to the supermarket – is not one made for good moral reasons. My contention is that, even where the relevant principals have not fallen short of a moral standard of care, and moral values have been included in the risk assessment, and the decision to delegate was morally neutral, they will be blameworthy for unforeseen harms if the decision to delegate did not serve a purpose that justified the risk from delegation. To quote from Hansson again,

> *"In order to appraise an action from a moral point of view, it is not sufficient to know the values and probabilities of its possible outcomes. We also need to know who exposes whom and with what intentions. To take just one examples, it makes a moral difference if it is my own life or that of somebody else that I risk in order to earn a fortune for myself."*
>
> (Hansson 2003, p. 302)

Jacinta and the self-driving car is a good example here. She was not, perhaps, particularly conscientious in her oversight, but she was not obviously falling below a moral standard of care towards road users and pedestrians, either. And there is nothing wrong, in and of itself, in taking pleasure in being driven around by one's self-driving car. But this decision did impose a risk upon the pedestrians on the way to the supermarket. And it imposed a risk of a particular kind of harm, as we have seen – a kind of harm different to that had Jacinta driven herself: one in which a person would be hurt directly by a machine, and in which a human may not have sufficient control over the event. Did Jacinta delegate to Myrtle for good moral reasons? Or did it just serve her convenience and fuel her delight in new technology? I explore such questions more below, in §4.5, when I answer the *locus of moral responsibility question* by applying the two stages of the procedure to the hypothetical examples given in the Introduction.

But we can state the conclusion of this discussion of responsibility as blameworthiness here first in terms of the three-term relation.

> *Even if a) principal S has not been negligent, and b) the commitment to risk incorporates risk assessment that responds to moral values, if c) the decision to delegate does not serve a purpose that justifies the imposition of risk from delegation to an autonomous system, then d) principal S bears a*

relation to the risk-exposed, in virtue of which S is blameworthy for all harms caused by the autonomous system, including UHC.

This is a morally demanding threshold for deserved moral blame. As such, even where the decision to delegate is morally neutral, the threshold for blame is lower than negligence. Imposing risk on others for the sake of one's own profit or gain constitutes failing to show the sort of interpersonal regard that we might expect from another and associated negative participant reactive attitudes, such as resentment, would be fitting.

Let us now consider the morally praiseworthy case. I will also state it in the terms of the three-term relation.

If a) principal S has not been negligent and b) the commitment to risk incorporates risk assessment that responds to moral values, and c) the decision to delegate does serve a purpose that justifies the imposition of risk from delegation to an autonomous system, then d) the principal S bears a relation to the risk-exposed in virtue of which S is not blameworthy for all harms caused by the autonomous system.

Many autonomous systems are deployed to achieve morally important goals. This might be the diagnosis of illness, or the distribution of vital supplies to hard-to-reach areas, for example. They can also relieve humans of tasks that physically or psychologically injurious to them, such as physically embodied systems that traverse across and clear a minefield ahead of human soldiers, or robots that clear up patches of nuclear waste instead of humans. And autonomous systems can carry out some tasks more accurately or reliably than humans can. Where the purpose is morally good, and the commitment to risk made for the right reasons, which includes recognition of the equity of the risk distribution implicitly accepted by the risk thresholds, this would, I submit, be a morally justifiable decision to deploy. Only in the case that the principals are morally praiseworthy in their decision to deploy, would they not be blameworthy for harms caused by an autonomous system. Anger and indignation would not be appropriate participant reactive attitudes towards them for unforeseen harms in these cases; pity or sympathy would be closer to the mark.[51] But responsibility still obtains even in

[51] This leads naturally to a question of moral luck, and specifically what Nagel calls "luck in the way one's actions and projects turn out" (1979, p. 28), and which has since been called 'resultant moral luck' by philosophers. Would it not be a matter of bad resultant moral luck for those who were deemed blameworthy on this morally demanding account, given that – for reasons beyond their control – many other principals who were equally non-praiseworthy in their decision at t1 'got away with it'? Though I do not engage in a treatment

blameless cases, in virtue of the fact that responsibility as attributability obtains. Gaps in blame do not entail gaps in responsibility.

4.5 Answer to the *locus of moral responsibility* question

Let me now state my answer to the central question in one piece. My project has been to use the framing of the delegation relation to answer the *locus of moral responsibility* question – which agents, S, are morally responsible for UHC, and in virtue of which properties, P – in two stages. First, the determination of the *loci of moral responsibility as attributability*. The normative asymmetry of the delegation relation yields a principal-proxy tracing principle for attributing moral responsibility for all harms caused by an autonomous system to the principal agents. This constitutes an exception to the necessity of their fulfilment of the Aristotelian conditions with respect to the immediate causal pathway to the harm. The delegation relation furnishes the P in virtue of which it is fair to attribute responsibility for UHC to principals S. The second stage is the determination of the *loci of moral responsibility as blameworthiness*. I have argued that, given the nature of the risk committed to as part of the decision to delegate, and the way in which this risk is imposed, principals S will deserve blame when the decision to delegate is not made for good moral reasons, which justify the imposition of the risk of harm upon moral patients.

Thus, the delegation framework can help us to structure responsibility interrogations for the outcomes and consequences of autonomous systems' behaviour in the real-world. And it yields a principle for attribution, alongside a heuristic principle for blameworthiness, that extends to all such consequences, including those which were non-negligently unforeseen. In addition to its descriptive accuracy and explanatory power, this is a normatively powerful result. It is morally demanding, both in attributing responsibility to all principals and in setting the threshold for blameworthiness far lower than the absence of negligence, or even including moral values within risk threshold assessments (even though both are necessary to

---

of moral luck in this thesis, I do not think moral luck is a fatal problem for the account. As Zimmerman argues, it is not that those who are now subject to moral appraisal are more blame*worthy*, or that the principals who had good resultant moral luck are *less* so. Rather, in the former case, their scope of blameworthiness is enlarged; they are open to blame for more things (Zimmerman, 2006, p. 599). Furthermore, as a practical point, by being non-negligent and incorporating moral values into risk assessments, even if the overall purpose of deployment does not adequately justify the imposition, a principal would be reducing her exposure to bad moral luck by taking this level of care and reducing the likelihood of harmful outcomes.

be excused from blame). It also demands that the decision is made for the reasons of sufficient moral strength that justify the imposition of the risk of harm upon people in the real-world operational domain.

It will be helpful to illustrate the actual working out of that delegation framework by applying it to one of the hypothetical examples from the Introduction. This endeavour needs be prefaced by highlighting another feature of my delegation account: it provides principled grounds for attributing moral responsibility and ascribing moral blameworthiness, but it still leaves room for case-based reasoning at the lower level, or in the more fine-grained discussions. Such reasoning would include considerations such as whether the purpose of delegation really *did* justify the imposition of harm, or whether the moral values *were* adequately capture in the risk assessment. In the discussion that follows, I do a little bit of this when speaking of the various actors and their relations, whilst framing the scenario as a whole according to the principal-proxy tracing principle and what I have called the 'heuristic' or rule-of-thumb for determining blameworthiness.

We can use the self-driving car case, in which the small child was injured in the road, to illustrate. We can identify the principals at each level of the complex decision to delegate to the autonomous system. The low-level principal is Jacinta, the sole owner and user of the self-driving car. Quite how realistic this is as a model of future car ownership is uncertain, given predictions of car sharing and more recourse to hired vehicles. Even so, we can imagine that Jacinta is a low-level principal of this kind, and is emblematic of other cases in which this principal is a personal owner of the system. Jacinta immediately hands over decision-making function to the system, to which she has given the nickname Myrtle. The high-level principals at the second level are the decision-makers in the regulators of the automotive industry and transport systems. These principals authorise the system as able to fulfil it agentive function in the real-world, on approved roads and highways, and within acceptable risk thresholds. The high-level principals at the third level are the vehicle's engineers. Principals at this level make sure that the vehicle fulfils its agentive function as an operational proxy for Jacinta as safely and effectively as possible. They do so by 'binding' the proxy in the operational domain. Even though Myrtle replaces Jacinta in the driving task, Myrtle's behaviour also represents the decision-making of the engineers, as guided and constrained by regulators. For this reason, I have said that autonomous vehicles are proxies with two kinds of principal, both low-level and high-level.

My claim is that, even though these principals are not directly causing the harm, the harm is directly morally attributable to them as if they were causing the harm personally. Each of these principals are the loci of moral responsibility as attributability. Though I have ruled non-reducible single group agents out of eligibility for moral responsibility ascriptions (§2.4.1), I take it that, if non-reducible groups *were* to have agency of a kind that warrants responsibility ascriptions – as, for example, List & Pettit (2011) and Pettit (2007) argue – what I say could apply to such groups, too.[52] But what I have in mind when I think of the principals, S, are individual and plural agents. I take plural agents to be reducible to individual morally responsible agents, who have shared obligations to one another and work towards a shared goal. While in practice it may be difficult to locate each principal, it is not impossible in principle, and a principle for doing so will be helpful for such real-world cases as they arise. That principle, I argue, is the principal-proxy tracing principle. To recall, this an exception to the necessity of instantiated Aristotelian conditions in the causal pathway to the harm. It therefore undermines a central premise in arguments about responsibility gaps.

The principal-proxy tracing principle is nonetheless complicated by the complexity of the multi-levelled decision to delegate to an autonomous system. At the heart of the tracing principle is the thought that, in virtue of instigating fulfilment of a delegated task to one's authorised proxy, one is responsible for all of the consequences of so doing as if one were acting personally, even though one is not literally acting personally. But it is not just a simple matter of Jacinta instigating this task by delegating to her trusted substitute driver Myrtle one sunny day. This decision of Jacinta's is dependent upon, and situated within, a deeper decision to deploy a Level 5 self-driving car. Those deeper decisions are made by high-level principals.

At each decision-point, these three features or criteria – instigation of the task, substitution of the principal, and authorisation of the proxy – are instantiated in subtly different ways and to different degrees. Jacinta meets the instigation, authorisation, and substitution criteria – but the authorisation criterion she meets is substantially different to and informed by the assurances given by high level principals at the second level. Jacinta does not much need to worry explicitly about the commitment to risk that her act of delegation involves, because

[52] Although this would be at the expense of a Standard View position as I have construed, which is restricted to human moral responsibility – and single group agents are not humans.

this has primarily been determined at the second level of the decision to deploy. Nonetheless she tacitly authorises the system as her proxy simply in virtue of delegating to it. High-level principals at the second level, the AV regulators, instantiate the authorisation criterion most strongly but only instantiate the substitution feature derivatively, in respect of the fact that they derivatively influence how it is a 'bound' as a substitute and they authorise it for the purpose of operational substitution. The high-level principals at the third decision point, the engineers who determine the final 'binding' of the system for deployment, meet the substitution feature insofar as how the system behaves in the world represents or reflects their decision-making about how it should best behave in the world, and fulfil its role as a substitute for a human operator. In addition, in validating the system as fit for deployment as it has been bound, these principals authorise that behaviour in the operational domain. This is where the overall decision to delegate gets particularly complex, but it is, nonetheless, an inherent, structural complexity in the case of delegation to autonomous systems: the system is a proxy with two kinds of principal: it substitutes the low-level principal in the fulfilment of the task, but how it does so is a representation of the high-level decision-making. I have expressed it thus: operationally, Myrtle is a proxy for Jacinta; but rationally, Myrtle is a proxy for the high-level principals.

I have said that the three internal criteria of the delegation relation: instigation, substitution, authorisation are instantiated in different ways by different principals in the complex decision to delegate. It might seem demanding to attribute responsibility to each of these decision-makers. But let us remind ourselves what responsibility as attributability is: it is the claim that the agent is open to moral appraisal and to demands of justification for the thing in question. And because there is more than one kind of principal, it is not the case that the high-level or low-level principal is solely or fully attributional responsible in these cases. It is the claim that they should answer for their part in the complex decision to delegate to an autonomous system that goes on directly to cause an unforeseen harm and hence for that harm itself. Given the moral gravity and significance of this decision, and the nature and form of the risk consequent upon the decision, it seems to be a normatively apt that principals should at least answer for their decision. The distance of the harmful consequence from the decision to delegate does not provide compelling grounds to evade or avoid these demands for justification. So, I stand by the bold claim moral responsibility as attributability traces to each of the principals. This makes my account a pluralist Standard View.

This means that each of the principals is open to moral appraisal for her part in the complex decision to delegate to the autonomous system. This openness to moral appraisal and to demands of justification is not itself responsibility as blameworthiness. Rather, the question of the principals' blameworthiness is determined by the answers and justifications they give. To recall, I have said that, aside from the obvious blameworthiness of negligence, a blameworthy decision to delegate will be one that disregards the moral features of the risk being committed to by the act of delegation (in addition to questions of intolerable severity and likelihood of harm), and which does not serve a purpose that justifies the imposition of the risk of the harm. I justified the moral stringency of this heuristic rule on the grounds that the decision to delegate to an autonomous system, which is also to commit to the risk of so doing, involves a particularly serious kind of risk – of harms caused by machines, which may be surprising, over which direct control and direct epistemic access will be lacking, and indirect guidance control may well be weakened and regulative control not meaningful, and which may not have occurred had a human been performing the same task.

We can now use the example of Jacinta and the self-driving car to break down the blameworthiness of decisions at each of the three key decision points. At the first level, Jacinta's decision to delegate the dynamic driving task to Myrtle, and thereby to commit to the risk of doing so, does not seem to betray any negligence on her part. It is within the operational domain, and the use and context are ones for which the vehicle has been approved as safe. We might think there is negligence in her superficial oversight (thereby giving the system substantial independence on the y-axis), and this will be a matter of case-based reasoning, but given the psychological and physical limitations of humans-on-the-loop, even her superficial oversight may be within tolerable bounds. Although her attitudes towards imposing risk on others will be implicit and manifest in the very fact that she does so, it is unlikely that Jacinta explicitly considers the moral features of the risk imposition when driving to the supermarket, but, as I have said, she does not need to worry about this too much – for it is something for which the high-level principals at the second level are answerable, primarily. So, does Jacinta's decision to delegate the task of driving to the supermarket to Myrtle serve a purpose that justifies imposing the risk it does on pedestrians along the route, such as the mother with her three small children on the pavement by the road? Again, it is a matter for case-based reasoning. It depends. If it is just a vanity trip in a shiny new toy, then no. But Jacinta may also have chosen the vehicle to reduce pollution. Perhaps she is not the sole owner, but a member of a carpool whose principal purpose for

delegating to a self-driving vehicle– as opposed to using a different vehicle, or walking, cycling, or taking the bus – brings benefits that override the risk imposition. Perhaps Jacinta chooses to delegate to the vehicle for reasons of safety. Whether or not those benefits override the risk will, again, be a matter of local level judgement. But the crucial question is whether it serves some morally good purpose in the best way possible. If not, given the stringency of my heuristic for determining blame, the injury to the child is one for which Jacinta is blameworthy, to some degree and not exclusively.

At the second level, the high-level principals' decision to authorise self-driving cars for delegation in the operational domain, and thereby to commit to the risk of this deeper dimension of the decision to deploy, does not seem obviously negligent either. But the hypothetical example has not been described in much detail. Providing some more detail will help us to see what sort of considerations would indicate negligence here. Negligence would include deliberately setting inaccurate risk thresholds, or accepting higher risk from an autonomous system in its intended context than is warranted. But, in addition, I have included as a consideration salient to blameworthiness whether moral values, such as whether the risk-exposed benefits from the delegation to the system. Perhaps, in this case, the benefits to residents of streets in which self-driving cars are allowed to travel is reduced pollution and fewer traffic accidents because, despite there being some accidents, the systems are far, far safer overall. Other considerations would include whether residents could avoid the risk if they wanted to, or whether it would become completely involuntary on their part. And finally, even if this has been considered but the decision to authorise is made, for example, to demonstrate the UK's technological prowess to the rest of the world, rather than to reduce fatalities and cut pollution, then the high-level principals would be blameworthy for the injury to the child.

At the third level, the high-level principals' decision to bind the system in a certain way, and thereby to commit to the risk of the consequences of this in the operational domain, may have been negligent at the point of validating the car's sensors. Dappled sunlight is not a rare phenomenon; it should have been detected during hazard analysis. If so, then these principals are blameworthy in virtue of that negligence. But imagine that it was not negligence at play here. Perhaps there was some particularly anomalous feature of this dappled sunlight in the real-world that undermined that validation of the system, which had been rigorous. Moreover, the safety-mitigation strategies from that point reveal respect for standards of

care. At this point, we should ask whether the binding decisions included evaluations of the moral dimension of the risk imposed. If there is a known inaccuracy in the feature, is it the pedestrians primarily who are exposed to the risk? And is this mitigated, perhaps, by the benefits to them? In addition, I have said, it is only if the purpose is a good one that justifies the imposition, such as to reduce fatalities and improve air quality, or deliver vital supplies to vulnerable individuals, perhaps, rather than to help the manufacturer sell as many cheap units as possible, that the injury will not be one for which the high-level principals at the third level are also blameworthy, to some degree.

Thus, my heuristic for determining blameworthiness is also morally demanding. Each principal is blameworthy for the consequences of their decision at each decision-point where this decision is negligent, disregards the moral nature of the risk, or is not made for reasons that justify the risk imposition. And this strictness or moral demandingness is fair, given the nature of the risk imposition from delegation to the system, as described above. It is interesting that, while the instrumentalist framework struggles to deal with unforeseen harms caused by autonomous systems non-arbitrarily, the delegation framework provides grounds to deal with them not only non-arbitrarily but in a morally exacting way.

This concludes my response to the *locus of moral responsibility* question in answer to the challenge from the suspension of the Aristotelian conditions. From the delegation relation, I have derived a principle for attributing moral responsibility to all principals in the complex decision to delegate to an autonomous system, and I have constructed a heuristic principle for determining blameworthiness. My argument defends a pluralist, Standard View position. Several kinds of human principal are morally responsible and there are no responsibility gaps; that is, their responsibility coverage is complete. But the Standard View also has an exclusivity requirement, which I address in the next Chapter.

Summary of Chapter:

In this Chapter, I have addressed the problem apparently posed to human moral responsibility by the human suspension of the Aristotelian conditions with respect to the immediate causing of the harm for which responsibility is sought. I have argued that the relation between human agents and autonomous systems is best understood as a relation of delegation. The delegation relation, or the principal-proxy relation, yields a

tracing principle that constitutes an exception to this suspension of ordinarily necessary conditions for moral responsibility. It enables us to trace the locus of moral responsibility as attributability for the unforeseen harm back to the low-level and high-level principals, each of which are involved in the complex, multi-levelled decision to delegate to the system. From this, we can determine the locus of moral responsibility as blameworthiness. These principals are also blameworthy for the unforeseen harm insofar as their decision within the multi-levelled complex decision to delegate, which is also a decision to commit to the risk of delegation to the system, is negligent, does not include moral values within the assessment of risk, or does not serve a purpose that justifies imposing the risk. This provides principled grounds for ascribing moral responsibility to humans for all harms caused by autonomous systems, include unforeseen harms, which still allows for case-based reasoning at the lower level.

CHAPTER 5
THINKING ABOUT THE FUTURE: A SOFT, PLURALIST STANDARD VIEW

I have defended a pluralist Standard View position. There are no responsibility gaps, because the principal-proxy tracing principle provides an exception to the necessity of the fulfilment of the Aristotelian conditions. But there is another requirement to the Standard View position, as I have characterised it – namely, that moral responsibility is borne exclusively by human agents (in our case, now, by the human principals). That is, it is also part of the Standard View that autonomous systems are not morally responsible for the harms they cause. In this Chapter, I interrogate that assumption, and consider what would have to be true for us fairly to ascribe moral responsibility to an autonomous system, and whether this is a conceptually coherent and physically possible proposition. I argue that it is.

5.1 When do proxies share moral responsibility with principals?

It has been my working assumption until this point that autonomous systems do not meet the first, exempting condition of moral responsibility: agential capacity. I have been officially agnostic, though grammatically cautious, on whether they are even intentional agents. I posit that the delegation framework remains true even though autonomous systems sit in indeterminate conceptual space, where they have goal-directed, causal agency but not intentional agency, and in which their autonomy is completely negative – defined in terms of independence from direct human control rather than the capacity and condition to set and realise their own ends. I have also argued that systems conceptualised in this way are still most accurately described as the delegates, or proxies, of human agents.

I agree with Di Nucci that it is senseless to say that moral responsibility is delegated to any proxy. Moral responsibility is something we possess or deserve; it is not something we transfer. It would be a category error to think otherwise (2020, p. 183-184). But it is also true that being delegated to confers new responsibilities and burdens upon one, and it opens new avenues in which questions of moral responsibility may arise. Of course, if autonomous systems fail the first condition of moral responsibility, it does not matter if delegation would open up questions of moral responsibility for a human in the same role. My purpose in this Chapter, however, is to scrutinise that assumption, and the grounds on which it rests.

First, given that the delegation framework remains in place for the framing of this question, it is instructive to consider human-to-human cases of delegation in which moral responsibility would not be borne exclusively by the principals. Kamm (2008) discusses the case known as 'Jim and the Indians', devised by Williams as an objection to the doctrine of negative responsibility – responsibility for the good outcomes one fails to bring about, which seems to be a commitment of utilitarianism (Williams, 1973, p. 98). Kamm makes two claims, and McMahan's response to the second claim expresses why it is that sometimes the case that proxies also bear moral responsibility for the consequences of an action or task that has been delegated to them. Kamm's first claim is that positive responsibility for the consequences traces completely to the principal in virtue of the principal's perpetration of the act (2008, p. 311). Her second claim is that the principal's full moral responsibility can make a difference to the permissibility or impermissibility of the proxy's action (2008, p. 312). Imagine a lawyer evicts a tenant for her client. The tenant is poor, and will be made homeless by the eviction, and the lawyer knows this. According to Kamm, moral responsibility for the negative consequences of the eviction, such as the tenant's homelessness, rests entirely with the landlord, who is the principal. Kamm uses the term 'agent' for my 'proxy'. Moreover, Kamm claims that "it can sometimes be morally permissible to be an Agent and do an act as an Agent, without its being morally correct for anyone (including the client) to do the act if one is not an Agent" (2008, p. 312).

With respect to this second claim, McMahan notes that this is a commonly held belief, of which people often persuade themselves, for example in military contexts*: "I was 'just obeying orders"* (2010, p. 678). But the locus of moral responsibility for an action is irrelevant to one's deliberations over its permissibility: it is not, or should not be, one of the reasons that counts either in its favour or against it. To return to the eviction case, the fact that the requirements of the lawyer's role may excuse her of blame for the negative consequences of the eviction does not render the eviction itself morally permissible. Moreover, the principal has what McMahan calls a "claim-right" (a right not to be prevented from evicting a tenant), which does not transfer to the lawyer, but the principal does not have a "liberty-right" (a permission to do wrong). As such, there is no right that the principal can transfer to the lawyer that would render the lawyer's action permissible. If the only positive reason the lawyer has for evicting the tenant derives from the lawyer's contractual duty to her client, the principal, "this duty may not be sufficiently strong to override the reason not to evict the tenant; nor is it

binding, for the lawyer can resign." (2010, p. 680). And I would add, though McMahan does not, that moral responsibility for agreeing to, or refusing to, undertake a morally impermissible action rests with the lawyer.

This objection has an implication for the allocation of moral responsibility in delegation cases. Kamm has argued that in those cases moral responsibility rests completely with the principal. McMahan's objection points to the fact that it does not, necessarily, rest completely with the principal — at least not in human-to-human delegation cases. Rather, if a principal agent delegates a morally impermissible or morally unconscionable task to a proxy, who duly carries it out, that proxy is not absolved of moral responsibility for its consequences. As a moral agent, she should not have done it, or should have overriding recourse to be exculpated on the grounds of duress or similar. In such a case, both the principal and the proxy are morally responsible for the outcome. Here, then, is one type of case in which moral responsibility is shared between principal and proxy:

> (Type 1) responsibility for consequences of morally impermissible actions that the proxy performs, knowing that they are impermissible and having the liberty to refuse to perform them. That is, 'I was just acting on orders' is not an exculpation when the order is to carry out a morally impermissible task (and the action is free).

On the assumption that the task is impermissible precisely because it is harm-causing, or highly likely to be harm-causing, Type 1 cases would concern foreseen harms. It would be a Row 1 case (see §2.1), and outside of our primary focus on the unforeseen harm. However, if a free proxy did knowingly carry out a morally unconscionable task, as the lawyer in Kamm's example did, then blame for the consequences rests straightforwardly with the principal and the proxy, as moral agents in their own rights. From the perspective of the principal, this would constitute one of the clearly blameworthy commitments to risk discussed in the previous Chapter, in which the very purpose the autonomous system serves is nefarious.

There is a second type case in which moral responsibility for a harmful consequence is shared between principal and proxy in human-to-human delegation cases. These are within scope, since there is no reason to think either the principal or the proxy would foresee the harms

so caused. This second type of case is when, even though the principal delegates, the proxy deliberately acts outside her mandate:

> (Type 2) responsibility for consequences of actions that the proxy deliberately performs outside of mandate, broadly construed (where this can include the adverbial - *how* the proxy agent performs the action - and not just the substantive - *what* action the proxy agent performs). That is, 'I was just acting on orders' is not an exculpation when one is not *just* acting on orders.

Within the parameters of the delegation framework, we could apply this human-to-human analysis to our case of human-to-machine delegation. We could say that, *if* autonomous systems met the first, agential capacity condition of moral responsibility, which they no doubt would meet if they were able deliberately to perform actions (as opposed to non-intentionally compute and implement outputs), and *if* the system deliberately acted against the principal's intentions, then responsibility for the consequences would rest with both the human principals and the autonomous system itself. It is a tall order for an autonomous system to meet this first condition, but my argument in this Chapter is that it is neither conceptually incoherent nor physically impossible, even though it does require some conceptual revision, on the one hand, and some quite considerable technical progress, on the other. The implication of this for my overall argument is that it makes my defence of the Standard View in this thesis what I have called a 'soft' rather than 'hard' one, in that it holds that is contingently and not necessarily true that moral responsibility rests exclusively with the human principals.

Before turning to the question of an autonomous system's meeting the first agential condition of moral responsibility, I should like to address the possible objection that it would follow that, should such an autonomous system that meets this condition ever exist, the moral responsibility for the consequences of its actions (not outputs) would be directly attributable to it. The system would be the *locus of moral responsibility* and the delegation framework would be redundant. But this objection, I think, goes too far. Certainly, it would make responsibility attributions to autonomous systems both possible and right in some cases. But for as long as the autonomous system is a proxy, it would go too far to say that it would bear sole moral responsibility. The asymmetries of the delegation relation remain salient to attributions of moral responsibility, just as they do to responsibility attributions in

human-to-human delegation cases. They are not just salient because of the apparent thinness of the autonomous system's agency, its lack of intentionality, or its lack of anything approximating positive autonomy. But instrumental frameworks for ascribing moral responsibility to toolmakers and tool users, if they are not to be essentially arbitrary, depend on the absence of such features.

5.2 Can autonomous systems be moral agents?

I approach the question of whether autonomous systems could ever be morally responsible agents by way of the question of their moral agency. In my discussion of the agential capacity condition in Chapter 2 (see §2.4.1), I said that moral agency is a precondition of morally responsible agency, but not sufficient for it. But we can start here, with the question of an autonomous system's moral agency. The possible phenomenon of artificial moral agency has been much discussed with that branch of applied ethics called 'machine ethics' (Anderson, Anderson & Armen, 2004). In this section, I consider some of those arguments and put forward a proposition of my own.

To frame this question, take the malarial Anapholes mosquito, or Frankfurt's spider moving across the table (1988, p.78). Imagine that it directly harms someone through its action of sucking blood or of biting. There is something goal-directed and minimally intentional about its doing so, and a mosquito or a spider is, plausibly, minimally sentient. In this sense, it seems to have greater agential capacities than what I have assumed is true of autonomous systems. Even so, we would not call the mosquito a moral agent. Why not? It seems ludicrous to say that the mosquito is acting morally (or, rather, immorally) when it transmits malaria to a sleeping person. It is not acting for moral reasons, to bring harm to a moral patient. I take it as a constraint that whatever conception of moral agent is determined, mosquitos should not fall into that category. I call this the 'mosquito test'.

One practical response to the prospect of autonomous systems causing morally significant harm independently of direct human control has been to build 'explicit ethical reasoners' (Dennis & Fisher, 2018; Cave *et al.* 2018), which are sometimes also called 'explicit ethical agents' (Moor 2006) or 'artificial moral agents' (Allen, Varner & Zinser, 2000). The idea is that these practical engineering projects will help to ensure that the actions taken by

autonomous systems will align with our ethical values and norms when they are operating independently of direct human control. Numerous ethical theories have been taken as inspiration by computer scientists and ethicists working together in this research area. To date, these have included: utilitarian theories and the representation of different ethical values as utility functions (Grau, 2006; Anderson, Anderson & Armen, 2004); broader consequentialist approaches (Winfield, Blum & Liu, 2014); deontic logic (Bringsjord, Arkoudas & Bello, 2006); Kantian deontology (Powers, 2006); prima facie duties (Anderson, Anderson & Armen, 2005; Anderson, Anderson & Armen, 2004); case-based or analogical reasoning methods (Honarvar & Ghasem-Aghaee 2009; McLaren, 2005; McLaren, 2003); forms of virtue ethics and "active learning" (Muntean & Howard, 2014; Coleman, 2001; Gips, 1995); and combinations of these approaches (Allen & Wallach, 2012; Wallach, Franklin & Allen, 2010; Wallach & Allen, 2009; Allen, Varner & Zinser, 2005; Allen, Smit & Wallach, 2000). At the time of writing, all such systems are in prototype phase. Explicit ethical reasoning systems may well be less harmful to humans than autonomous systems that have only implicitly, if at all, been developed with ethical values. But our question is whether they would be moral agents, and therefore at least in the ballpark for responsibility for any harms they do cause. Sometimes, the following claim is made of the machines of the practical machine ethics project: since these devices take morally correct actions, they merit the ascription of moral agency (Anderson & Anderson, 2007, p. 19). Those who argue that machines are not, nor cannot be, moral agents tend to point to some feature that is essential to moral agency that machines necessarily lack. The missing features are identified variously as original intentionality (Johnson, 2006), understanding (Stahl, 2006), the capacity to suffer (Sparrow, 2007), and the capacity to act for the right reasons (Purves, Jenkins & Strawser, 2015). Those who argue that there is conceptual space for artificial morality either maintain this missing feature is not in fact necessary (Wallach & Allen, 2009; Floridi & Sanders, 2004) or argue that this feature is not in fact missing, and that some machines possess it (Powers, 2013). My claim is that the requisite features are almost certainly missing now, but might not always be. This grounds my soft Standard View.

The question of artificial moral agency can be disambiguated in the following way:

> Sense-1: Autonomous systems as morally relevant agents, capable of causing harm to a member of the moral community;

Sense-2: Autonomous systems as moral agents, capable of acting morally;

Sense-3: Autonomous systems as participants in a wider, multi-agent moral system, in which distributed actions can cause harm to members of the moral community.

Autonomous systems can already accurately be described (so long as we are happy to speak in terms of their being agents, in the causal and goal-directed sense) as agents in sense-1 and sense-3. What is less clear, but also necessary if one is to be eligible for an ascription of moral responsibility, is whether autonomous systems are, or can be, agents in sense-2. The mosquito is a morally relevant agent in sense-1. But the mosquito does not act for moral reasons. In this section, I work through some of the key arguments for artificial moral agency within philosophical machine ethics, and consider the gradations by which they develop, and whether any get us to a machine that is acting morally, which would be an artificial moral agent in sense-2.[53] If so, then we *may* have an autonomous system with the sort of agential capacity that would qualify them to bear moral responsibility – and therefore, in the presence of the right conditions and circumstances (such as a Type-2 case), to share moral responsibility with their human principals.

The reader might query my switch to speaking of autonomous systems as agents, which I have avoided doing until now. But the question of agency is now unavoidable. As discussed in Chapter 1 (§1.1), the technical sense of 'agent' is well established, and what I have called 'autonomous systems' could almost entirely have been called 'artificial agents' with no real difference. Artificial agents are interactive systems, embedded in environments. Artificial agents are also goal-directed systems. They are programmed with objective functions, reward functions, or fitness functions, for example, to which optimisation of their functioning to our goals is aligned. In addition, they are clearly causal agents: they have causal powers, they can effect changes on other things.

It is an open question as to whether they are intentional agents, able to perform actions that are intentional under at least one description (Davidson, 1963; Anscombe, 1957). The standard conception of such agency is that to act intentionally is to act for a reason, which is

[53] Fossa (2018) and Torrance (2011) draw a useful distinction between 'practical' and 'philosophical' machine ethics. Practical machine ethics is the engineering project of actually building explicit ethical reasoners. Philosophical machine ethics, under which aegis this discussion sits, is focused on such questions as whether advanced technological systems can be moral agents, bear moral responsibility, have rights and moral standing, and whether our conceptual schemes can accommodate them in this way.

to act in a way that can be rationalized by reference to the agent's goals or intentions, beliefs, and desires (Schlosser, 2019). Such a capacity is necessary therefore, if one is to act morally, or for moral reasons. Intentions, beliefs, and desires are generally called 'intentional states'. Computational systems, particularly drawing on Bratman's (1987) work on practical reasoning, are certainly designed with *representations* of intentional states. Yet this is not to say that the systems are literally intentional, and that they have beliefs, desires, or intentions of their own. A common dialectical move, particularly when faced with computational entities of uncertain or indeterminate properties, is to tackle the question side-on, obliquely – not to address the instantiation of those properties directly, but to see if something else is true that would imply their possession or make the ascription undeniable. Turing (1950) takes this approach in his construction of the 'imitation game' or Turing Test, in which an interrogator asks questions of human respondents and a computer, in an effort to discover which is which. As Millican puts it,

> *"… if the computer were able to give sufficiently humanlike responses to resist identification in such circumstances, then it would be quite gratuitous to deny that it was behaving intelligently, irrespective of its alleged lack of a soul, and inner perspective, consciousness, or whatever."*
>
> (Millican, 1996, p. 2)

Dennett (1989, 1978) can be seen as taking a similar kind of approach with the Intentional Stance. If we can best explain and predict a machine's behaviours through the ascription of intentionality, we should ascribe intentionality to it. Some of the arguments I now survey take similar approaches, as does my own.

5.2.1 Source-of-moral-actions argument

The first main argument for artificial moral agency I consider comes from Floridi & Sanders' influential 2004 paper, 'On the morality of artificial agents'. The authors argue that the standard conception of a moral agent fails to account for autonomous systems that are able to cause harm to humans "independently of the humans who created them" and that this "hinders the development of a satisfactory investigation of distributed morality" (2004, p. 351), in which distributed actions can cause harm to members of the moral community (sense-3). In later papers, Floridi (2016, 2013) develops this notion of distributed actions, and of distributed moral actions, as part of a wider project to show that our "standard

perspectives [...] run the risk of remaining unduly constrained by an anthropocentric conception of agency" (2013, p.728). It is the question of agency (sense-2) that is our current concern.

To derive the conclusion that artificial agents can be moral agents, Floridi & Sanders deploy a methodology by which a selection of observable features of moral agency are abstracted from all other features, and the entity is characterised by observably instantiating those abstracted features. Floridi & Sanders argue that artificial agents can be moral agents at a certain 'Level of Abstraction', which is the framework for this methodology. When such agents cause moral harm, they are "legitimate sources of im/moral actions" and the set of moral agents should therefore be expanded to include them (2004, p. 352). I call this the Source-of-Moral-Actions argument.

Despite the influence of this paper, it is not often reconstructed in such a way as to reveal its structure and inferences.[54] I give a systematic presentation of the argument here:

> 1. Whether or not a particular entity qualifies as a moral agent depends, in the first instance, on the specific Level of Abstraction (LoA) at which one chooses to analyse it.
>
> 2. The correct LoA at which to analyse an entity for the sort of agency that grounds moral agency includes three observable features: interactivity (the entity and its environment act upon each other); autonomy (the entity can change state without direct response to interaction); adaptability (the entity's interactions can change the transition rules by which it changes state).
>
> 3. To be a moral agent, this entity must also be capable of 'morally qualifiable action' – where morally qualifiable action is action that 'can cause moral good and evil'.
>
> 4. An agent that satisfies 2 and 3 is a source of morally qualifiable action.
>
> 5. A source of morally qualifiable action qualifies as a moral agent.

[54] Behdadi & Munthe (2020) and Burr, Christianini & Ladyman (2018) are exceptions to this trend.

6. Certain artificial entities do in fact satisfy 2 and 3.

C. Therefore, those artificial entities (at 6) qualify as moral agents (from 4 and 5) and the set of moral agents should be expanded to include them.

According to the authors, examples of artificial entities that instantiate the three observable features (premise 2), and which are capable of morally qualifiable action (premise 3), include: futuristic thermostats (which monitor not just ambient temperature but also, for example, a hospital patient's state of well-being); web-bots (such as those that filter email messages, which is a morally relevant task because "we value our email" (2004, p. 370)); certain machine learning systems (specifically, those with outputs that can benefit or harm humans); and organizations. This being so, according to the argument, these qualify as moral agents. I think this argument fails the mosquito test. There is a Level of Abstraction at which the Anopheles mosquito is interactive, autonomous, and adaptive (premise 2). And it is clear that this mosquito is the source of actions that can cause grave harm to moral patients (premise 3). By the argument's lights, then, the Anopheles mosquito is a moral agent. But the mosquito is not acting immorally when she bites and infects a sleeping child. The mosquito is an agent in sense-1, but not in sense-2.

As an aside, Floridi & Sanders' argument also equivocates on the notion of 'sourcehood' of moral action - between the system's being a *causal* source of the action and its being a *moral* source of the action. There is an elision in the argument from the claim they are causal sources of morally qualifiable action (2004, p. 351) to the claim that they are moral sources of such actions (2004, p. 367 & p. 376) – that, for example, these actions issue from their own practical reasoning. This inference from causal to moral agency is not warranted.[55] And yet it is far from innocuous. Only the latter sense justifies the assertion at premise 5. But it is not itself justified by the conceptual analysis at premise 2.

Floridi & Sanders' argument yields merely agency in sense-1, not agency in the crucial sense-2. Their secondary purpose is to account for the involvement of autonomous systems in our distributed and interconnected moral systems (sense-3). It is not clear to me how elevating

[55] Floridi might press back against this view. In his later papers discussing distributed moral actions arising out of interactions at the local level between individual agents (2016, 2013), Floridi argues in favour of assimilating moral responsibility to causal accountability: "... talking about 'responsibility' in the aetiological sense of being the source of (i.e. causally accountable for) a state of the system, and *therefore*, being morally answerable (blameable/praisable) for its state." (2016, p. 6) (italics mine).

morally relevant agency into moral agency is explanatorily helpful. It is perfectly possible to account for autonomous systems by acknowledging that they are morally relevant agents that can operate independently of direct human control. This would be enough to establish their moral significance, without imputing to them moral agency.

5.2.2 Causally-efficacious intentional states argument

A second argument for artificial moral agency comes from Powers' 2013 paper 'On the Moral Agency of Computers.'[56] I also include within this approach aspects of the arguments from Sullins (2011) and Himma (2009). Their common thread is that, to qualify as a moral agent, the artificial agents should have intentional states that cause its actions. Powers holds that the central question for the possibility of artificial moral agency is whether the actions of an autonomous system are "of the sort that qualify as moral actions as a result of the computer's decisions" (2013, p. 228). On the face of it, this is not dissimilar to the idea underlying the Source-of Moral Action argument, since both seek to establish the autonomous system as not just conduit of, but in some relevant sense the source of, moral actions. But while Floridi & Sanders' argument abstracts away intentionality as an unnecessary feature, this argument seeks to locate the autonomous system's intentionality, and from thence defend a philosophical claim about artificial moral agency.

Powers argues that computational agents can have "internal intentional states" which constitute their own reasons for action and, where those actions are morally relevant (sense-1), such agents are moral agents. A systematic presentation of his argument as it is given in the paper is as follows:

1. Autonomous systems can possess internal intentional states.

2. These internal intentional states constitute an autonomous system's reasons for acting.

3. An agent's reasons for acting cause that agent's actions.

[56] To note, this is not second chronologically, but in the development of the conditions a machine should meet to qualify as a moral agent.

4. When an agent's actions are caused by their reasons for acting, the agent is acting from its own reasons.

5. When autonomous systems acting from their own reasons act in ways that have morally relevant effects on moral patients, they are genuine moral agents.

C. Therefore, autonomous systems can be genuine moral agents.

Powers claims that premise 1 is supported by the computational theory of mind, and that premises 2 and 3 are supported by Davidson's (1963) account of action. Rather than engage with exegesis on these points of theoretical defence, let us accept for the sake of argument that the first three premises are true. The main question for the argument in this Chapter is not just whether intentional states constitute reasons for acting, but what the *content* of those reasons would be. After all, the intentional states themselves – say we think of them as intentional states in virtue of the causal and functional role they play in the input-to-output process – may not refer to the effects of the actions (or outputs) on moral patients or have any specifically other-regarding content; these may be merely side-effects.

Powers' argument includes intentionality, and provides a theoretical framework for its attribution. But it still fails the mosquito test. If internal intentional states constitute reasons for acting (and are the reasons from which we act), well, a mosquito has these, too. But the mosquito does not act from moral reasons. There is more to the story of artificial moral agency that we need to tell. As it stands, this argument also yields agency in sense-1 rather than sense-2, and therefore not the sort of moral agency required to ground an ascription of moral responsibility.

5.2.3 Functional morality argument

Another influential position in defence of artificial moral agency comes from the work of Wallach and Allen, and their collaborators.[57] In many ways, the authors' primary concern is the practical machine ethics project*:* "explicitly building ethically appropriate behaviour into autonomous systems" (Allen, Smit & Wallach, 2005, p. 149). But at times the project is also one that falls under philosophical machine ethics: "… we think that the notion of functional

[57] In addition to the papers cited, see also: Allen, Varner & Zinser, 2000; Wallach, Franklin & Allen, 2010.

morality for machines can be described philosophically and pursued as an engineering project" (Allen & Wallach, 2012, p. 62).

It is the philosophical description that concerns us here. Their philosophical claim is that the existence of autonomous systems programmed to instantiate functional morality "will expand the circle of moral agents beyond humans" (Wallach & Allen 2009, p. 4). Despite variations between the different publications, the core of this position is that suitably programmed autonomous systems can be described as artificial moral agents when located at a point on a continuum between 'operational morality' and 'full-blown moral agency'. This point is 'functional morality.' In their 2012 paper, Allen & Wallach say the most common objection to the argument of their 2009 book *Moral Machines* (Wallach & Allen, 2009) is that, because autonomous systems lack a certain property that is necessary to 'full moral agency,' it is a conceptual error to speak of artificial moral agency at all. Their response is that they accept, for the sake of argument, the antecedent (the lack of full moral agency) but deny the consequent (the conceptual incoherence of a claim of moral agency). For the authors, there are gradations of moral agency, there are "steps between" functional and full moral agency (2012, p. 62). As Fossa notes, moral agency as a scalar phenomenon implies that there is no essential difference between full or 'genuine' moral agents and artificial moral agents (Fossa, 2018, p. 116): they are different in degree, or in their range and richness of capacity, but they are not (at least, *qua* moral agency) different in kind.

The argument can be expressed as follows:

> 1. An autonomous system can be programmed to take, monitor, and assess its actions for their effects on moral patients.
>
> 2. Such an autonomous system therefore acts as if it is a moral agent.
>
> 3. An autonomous system that acts as if it is a moral agent is a functional moral agent.
>
> 4. The standard set of moral agents should be expanded to include functional moral agents.

C. Therefore, the standard set of moral agents should be extended to include any autonomous systems that are programmed as at 1.

The 'as if' of premise 2 can be read in two ways. The first is 'as if' as in simulation. It leads to a simulacra view of artificial moral agency. Naturally, a simulacra view cannot ground moral agency in sense-2. The second is 'as if' as in equivalent structure of behaviour. This possibly can ground moral agency in sense-2. The question is whether equivalent structure of behaviour is implied by the first premise, such that the autonomous system's taking, monitoring, and assessing its action (or outputs) does play the same causal and functional role in the system's behaviour as it does in agents we take, uncontroversially, to be moral agents.

A natural objection at this point is that the system is *programmed* to behave in this way, and so its behaviour is not structurally equivalent. This sort of objection might be levelled by those who argue that an autonomous machine cannot act for moral reasons because it is not self-motivated to act, rather it "simply manifests an automated response which is entirely determined by the list of rules that it is programmed to follow" (Purves, Jenkins & Strawser, 2015, p. 861). This objection assumes a conception of autonomous machines that is closer to the rule-based automata described in Chapter 1. With adaptive learning systems, with high degrees of autonomy on the z-axis, high-level designers and engineers under-determine the system's responses, but do provide it with the internal capacity to adapt to is environment. This *could* ground the functional equivalence of behaviour at premise 1. A great deal depends on how that premise is understood, and how structurally equivalent the taking, monitoring, assessing (and acting) is required to be, and whether this constitutes acting morally.

In the next sub-section, I propose a different form of functional morality that goes deeper than that posited by Wallach and Allen. But it should be highlighted that this argument *does* pass the mosquito test because mosquitos flatly fail the first premise. If it is a constraint that whatever conception of moral agent is determined, mosquitos should not fall into that category, this conception operates within that constraint.

Another approach is to take a more relational view. This is the perspective taken in Coeckelbergh's quite different 'as if' argument for artificial moral agency: a defence of virtual moral agency (2009). His argument is specifically focused on humanoid robots, but this

morphological feature can be abstracted out of the discussion without disservice to the central claim. Coeckelbergh argues that we ascribe moral agency to others on the basis of "how they appear to us and what we experience when we interact with them" (2009, p. 184). We take an 'as if' approach to moral agency in our everyday lives all the time. Future autonomous systems may well appear to us and be experienced by us in ways that also warrant this ascription. Coeckelbergh calls this 'virtual moral agency'. We do not experience mosquitos in this way in our interactions with them. Coeckelbergh, I think rightly, sets the bar high as to what would count as sufficiently compelling appearances of virtual moral agency. I draw on his insight about experiences with, and relations to, the system in my own proposed argument for artificial moral agency in sense-2 below.

Floridi & Sanders, who began our current wave of the discussion of artificial moral agency, touch an important central nerve in pointing out that our conceptual frameworks are ill-equipped to deal with independent, interactive, and adaptive machines capable of causing harm to members of the moral community, and to account for their role in our wider moral system (sense-3). But the purposes of such a conceptual framework are not just descriptive, they should also help us in the question of ascriptions of moral responsibility. Even if moral agency is not a sufficient condition of moral responsibility, on my account, it should still be of the right kind that would ground moral responsibility.

The first two arguments failed to do this. We discovered from Powers' argument that the issue is not missing intentionality but missing moral reasons, without which autonomous systems fail the mosquito test. The functional account, understood as a structural equivalence of a kind of moral behaviour, is more promising for the legitimate predication of moral agency to autonomous systems. It avoids the mosquito *reductio*: mosquitos, unlike such systems, do not evaluate the effects of their behaviour on moral patients. My project in this Chapter is to bolster these accounts by introducing a notion of acting morally that I think that autonomous systems literally could instantiate: altruistic behaviour.

The implication of this for my overall argument is that, if such systems were deliberately to deviate from their task or mandate, as per a Type-2 case as described at the start of this Chapter, then they could share moral responsibility with their human principals. The delegation framework can account for this perfectly well, but it does mean that the exclusivity

claim of the Standard View would not be upheld. As such, my argument becomes a soft, Standard View position.

5.3. Argument from other-regarding function

I have described the key difference between sense-1 (morally relevant agency) and sense-2 (moral agency) as that the latter involves acting morally, which I have cashed out primarily as acting for moral reasons. One way to address the difficult question of whether an autonomous system could rightly be said to act for reasons (and, specifically, for moral reasons) could be whether something else, more discernible, could be true which would imply this. I have mentioned that Turing's Imitation Game and Dennett's Intentional Stance arguments do this in different ways. My proposition is that one capacity that would *imply* their acting from moral reasons, even if only minimally, is the instantiation of genuinely altruistic behaviour, in which the systems subordinated the fulfilment of their own goals to help a moral patient meet her morally important goals.

A thin understanding of altruism would simply involve doing something in someone else's interests. This is captured in the functional morality accounts of the preceding sub-section, in the emphasis on the system monitoring the effects of an autonomous system's behaviour on moral patients. But a deeper account of altruism recognises that there can be something personally costly to the altruistic agent in bringing about good effects to others. An altruistic agent overrides meeting her goals to respond to the needs, interests, or goals of others. This, I argue, is another form of functional morality – which autonomous systems could instantiate, and which would furnish moral agency in sense-2.

My argument is as follows:

> 1. An autonomous system has a goal, or set of goals.
>
> 2. This autonomous system could learn in its environment to subordinate the fulfilment of its goal to help a moral patient meet her (perceived) goal.
>
> 3. If, in the presence of the right conditions, an autonomous system were to do this, it would be acting morally.

4. An agent that can act morally is a moral agent.

C. Therefore, if an autonomous system were to act as at 3, it would be an artificial moral agent.

If an autonomous system were to do this, we would *have* to infer that its agency is intentional. I think this also the passes the mosquito test. At least, if a mosquito could learn to put others first in this way, I would be happy to call the mosquito a moral agent. I should add that my argument – which is aimed at showing the Standard View claim of exclusivity of human moral responsibility is not necessarily true – does not require that autonomous systems of this kind are technically feasible, or even possible on today's technological capacities. But it does require that such systems are at least physically possible, and that, should they come about, it would conceptually legitimate or coherent to think of them as moral agents (sense-2), even if it is admitted that they would be thin and bloodless versions of a kind.

Below, I discuss each premise of the argument in turn, with the exception of premise 4.

1. An autonomous system has a goal, or set of goals (e.g. reward function)

In Chapter 1, we saw that one defining characteristic of autonomous systems is that they are goal-directed entities. The systems are programmed with goals, which take the form of mathematically formalized functions. In learning-based systems these goals may be objective functions (when programmed with supervised learning algorithms), reward functions (when programmed with reinforcement learning algorithms), or fitness functions (when programmed with evolutionary algorithms). These programmed goals are aligned to the system's agentive function – the decision-making task that it fulfils.

To be clear, there are two different sets of goals in play. First, there is the mathematical function, such as the objective function or the reward function or the fitness function, which we can call goal_{A}, to show that it is the *autonomous system's goal*. Second, there is the specified agentive function, which is the goal that the human principals have for the system, such as driving autonomously without collision, which we can call goal_{H}, to show that it is the *humans' goal for the system*. Goal_{H} is our goal for the autonomous system, but it is not the system's goal.

$Goal_A$ is the system's goal. $Goal_A$ is aligned to $goal_H$, and it is causally related to $goal_H$, but it is not reducible to $goal_H$.

I anticipate two objections. I shall take each in turn.

The first is that $goal_A$ is not really a goal. This objection might take its purchase from the autonomous system's lack of conation: there is no striving or trying or willing to reach the goal. It is not aiming at a goal, the objection might go, it is running in a feedback loop. There are two responses to this. First, it is not clear that striving to reach something is either (logically) necessary or sufficient to that thing's being a goal. It is the necessity claim that concerns us here. One may be goal-directed without conation. Autonomous systems are intrinsically goal-directed entities. They are built that way. More precisely, they are built to optimize for $goal_A$.

This response may lead to the second, more likely objection that $goal_A$ is a goal, but it is not really the *autonomous system's* goal. The high-level principals set or implement the system's goals. Even so, a goal's being set by others need not mean that the goal does not pertain to the goal-possessing agent. There are many human goals, for example, that we do not set for ourselves. Our biological, existential goals are a part of our constitution; we do not set them; we do not choose our goals of eating, or otherwise staying alive. And less loftily, my employer might set my goal of making 100 widgets a day. The fact that this goal has been set for me, and I have been directed at achieving it, does not make it any less my goal. It *becomes* my goal.

To conclude, the autonomous system has a goal: $goal_A$. This is a purely quantitative goal. The system is designed to be a self-interested agent; its behaviour is directed at meeting, and optimizing for, $goal_A$ – and its doing so aligns with the human goal for the system, which is its specified agentive function.

2. This autonomous system could learn in its environment to subordinate the fulfilment of its goal to help a moral patient meet her (perceived) goal

My claim is that $goal_A$ is the reference point against which the system's behaviour can be identified as altruistic. If the system subordinates its fulfilment of its own $goal_A$ then this is sufficient for altruistic behaviour. To be clear, the system in question might behave

altruistically towards either a human agent or towards another system. I restrict my discussion here to behaving altruistically towards a human agent. Now of course the system could be explicitly trained to do this. Reinforcement learning, for example, includes delayed rewards which ultimately lead to greater optimisation of reward function. But a claim to altruistic action that constitutes acting morally (as at 3), which includes acting for moral reasons, would be better supported by learning it without its being explicitly trained or set as a performance criterion against which to test and validate the system. To 'pick it up' in the operational domain, through interaction with other agents, captures more of the grasping of moral reasons that we require.

There is, therefore, an adaptiveness requirement on the system, such that it continues to learn in the operating environment, and it must also have the capacity both to perceive or to detect, through its sensors, the goals and interests of humans and to evaluate whether those goals are being met. This requires that the system non-accidentally identify certain features of the environment, and accurately interpret transactions and relations between agents in the environment. Physically, this is not impossible, but technically and in practice it would be extremely hard. It would require a sophisticated sensitivity to human behaviour. To be clear, for my argument, all that is required is that an autonomous system's behaving in such a way is a physical possibility, not a technical possibility or even plausibility.

There are some pathways by which it might be achieved. One way this behaviour might develop is through detecting patterns of human behaviour in the operating environment in which people subordinate their own rational maximization of expected utility to enable the moral patient to increase theirs. Moreover, computational modelling shows that helpful agents do better. Danielson (1992) conducted a set of experiments with virtual robots in game theoretic scenarios. These experiments showed that co-operative, reciprocating software agents are more successful in Prisoner's Dilemma-type virtual games than un-cooperative competitors. To be clear, Danielson explicitly programmed some robots to behave in this way; it was not an adaptively learnt behaviour. And Danielson's purpose was to demonstrate the rational justification of morality (where 'rational' is to be understood in the self-interested sense given to 'rationality' in rational choice theory) rather than the possible moral agency of machines. Nonetheless, this experiment provides grounding for the technical possibility of altruistic behaviour for a goal-directed system in a social environment. Rational, self-interested, adaptively learning systems might calculate that falling short of $goal_A$

in the near term, by responding to other agents as moral patients, contributes to greater and more sustained optimization of $goal_A$ in the long-term.

Of course, the natural objection at this point is that this would not be altruistic behaviour at all, if in the end it is simply optimization for $goal_A$. But let us take stock. This is not immediate optimization, or a highly bounded delayed reward function, it is a sophisticated extrapolation of reasons from a living environment that results in a system's responding to moral patient's needs or goals at the expense of its own. This is surely sufficient to meet a minimal characterisation of altruistic action. As I mentioned in above, the fact that it contributes to the agent's own interests and goals does not stop it from being altruistic. Altruism and co-operation help humans to achieve their goals more consistently, too.

I now invite the reader to consider the following thought experiment to bring premise 2 into sharper focus:

> *Imagine that the self-driving car, Myrtle, in our first hypothetical example is carrying out the whole entire driving task without Jacinta's intervention and with minimal oversight. Many of its sub-systems (those components that interpret data from the sensors, and the path planning and decision-making components) are complex ML-models. The vehicle learns about and adapts to its environment over time. Myrtle detects in one of the bounded boxes in its visual field, that a person is lying down on the pavement. A pushchair is next to her, and three small children. Despite having calculated the safest path and trajectory, the vehicle deviates from the task Jacinta has transferred to it of driving to the supermarket and diverts from its optimal, safest trajectory – both of which are sub-optimal for its $goal_A$ – to stop by the person lying down on the pavement. Myrtle then calls an ambulance.*

This would be subordination of the fulfilment of its goal to help a moral patient meet hers. And I think this would be altruistic behaviour. My next claim is the premise that this would be enough for sense-2 moral agency.

3. If, in the presence of the right conditions, an autonomous system were to do this, it would be acting morally

The next claim is that altruistic behaviour of this kind *would* constitute acting morally. There is the fact that the system would be helping another at its own expense. I think this would imply, even if only minimally, acting for moral reasons. This in turn would imply the possession of at least rudimentary intentional states, including the intention, expressed in the action, to help the moral patient meet her need or goal.

Naturalistic accounts of morality can be adapted to support this view. In particular, Wong's meta-ethical theory of moral pluralism can be extended to this case (2009). Wong's "functional conception of morality" looks to biological and cultural evolutionary theory. Wong presents a universal, functional criteria account of morality as having two dimensions. The first is the interpersonal dimension: morality gives rise to beneficial social co-operation. The altruistic behaviour of the kind I have depicted could give rise to better reciprocal relations between human and autonomous systems, or to better relations within the moral system (sense-3) more generally. The second is the intrapersonal dimension: morality gives rise to individual flourishing. And as we have seen, the altruistic behaviour of the kind I have depicted optimises for the system's goals. In addition, Wong claims that, to fulfil this interpersonal and intrapersonal function adequately, a morality must address the basic propensities, interests, and needs of human nature (2009, p. 39-47). This too is instantiated in the thought experiment. Naturally, this involves some modifications to Wong's account. Wong is not in the least concerned with the moral agency of machines. Whereas he looks to the interpersonal and intrapersonal functions of morality as the universal criteria for a true or acceptable morality, I propose this is extended to inter-agential and intra-agential functions of morality. Another modification I make to Wong's account is to extend the basic interests, propensities, and needs of human nature, to include *goals*, including the system's goals. With these modifications, the proposition that autonomous systems could adapt or evolve to behave in ways that fulfil the universal, functional criteria of morality is not only conceivable, but physically possible. The reference of 'acting morally' would be whichever entities fulfil the universal, functional, other-regarding criteria of morality; this does not have to be an anthropocentric standard.

One might be inclined to develop the objection here that, even if there is a sense in which behaviour such as that described in thought experiment really is altruistic, this isn't really acting morally. The systems are merely adaptive, rational, goal-maximising agents. This line of thought is akin to the evolutionary debunking-style arguments that none of our apparently

moral practices are at root 'moral'. Evolutionary explanations of moral phenomena point to traits such as "reciprocal altruism" which can increase the agent's fitness or chances of survival and can deliver collateral benefits to kin members over time (Axelrod, 1984). An evolutionary debunker argues that evolutionary explanations of moral phenomena – what Joyce calls "plausible speculations" (Joyce, 2001, p. 135) – are complete explanations of moral phenomena, which obviate the need for moral truth (Joyce, 2006; Street 2006; Joyce, 2001).

To defend acting altruistically as acting morally against this reductive, debunking claim, we could accept that if 'moral agent' is simply reducible to 'adaptive, fitness-enhancing agent' then so be it. If human moral agency reduces to this too, it does no disservice to my argument. In any case, a plausible evolutionary explanation of behaviour does not entail that it is not *also* moral behaviour. In fact, the evolutionary account is consistent with moral realism. Copp argues that an evolutionary account on which we judge pro-social actions, such as helpfulness, to be morally good just in case they contribute to the needs of societies (continued existence, stable cooperation, peaceful and productive behaviour amongst its members), actually offers a vindication of our moral judgements (2008). Copp is concerned with a society's needs, because he defends a society-centred moral theory, but it does not seem a stretch to say that an evolutionary account might also vindicate our judgements that helpfulness is morally good just in case it contributes to the fulfilment of the basic needs of human agents.[58]

C. Therefore, if an autonomous system were to act as at 3, it would be an artificial moral agent

Were autonomous systems to start acting morally like this, we would be justified in thinking of them as artificial moral agents (sense-2) within a wider moral system (sense-3). Thus, I agree with the earlier arguments surveyed that the class of moral agents can be extended to autonomous systems. My purpose here has been to open up the conceptual space around the possibility of artificial moral agency. The position does not obscure the deep and significant differences between artificial and humans. Rather, the argument is that those differences do not preclude both artificial and humans from meeting the universal criteria for the form and function of morality, and so the account is conceptually coherent. Even so,

[58] The full statement of Copp's society-centred moral theory is given in his book *Morality, Normativity, and Society* (1995). But unlike his 2008 work, this does not discuss its connection to evolutionary theory.

given that the anthropomorphic standard has historically determined the reference of acting morally, or moral agency, this account does endorse some conceptual revision. The conceptual revision is not that we should modify the concept of moral agency to the degree that sense-1 is suddenly sufficient, and that entities that fail the mosquito test count as moral agents, as some of the arguments earlier discussed would have it. This would set the threshold for moral agency too low. Rather, it is that we should revise it to include all entities that can take other-regarding, altruistic action, helping others to some cost to themselves. This, I have speculated, could be construed as their acting for moral reasons. If so, then this would be one pathway by which autonomous systems *could* meet the first agential capacity condition for moral agency.

5.4 Would artificial moral agents also be morally responsible agents?

It is conceivable that a system might start to act for moral reasons (sense-2), in the way I have described, and not yet have deep rational control or awareness of what it is doing. I would suggest that small children, for example, can do genuinely kind things for others, comfort another in distress and set aside immediate pursuit of their own goals to do so. Some higher order animals could be construed as doing this, too. In other words, acting for moral reasons is not sufficient for morally responsible agency.

Would the artificial moral agents I have just suggested meet these deep conditions? What else would have to be true for this to be so? To situate this claim within the broader argument, if an autonomous system could meet this condition and were to act as in a Type-2 case – that is, directly take action outside of their mandated or delegated task, as the vehicle did in the case above with the fallen pedestrian on the pavement – then this would be a case in which the autonomous system would share moral responsibility with its human principals. The principals would still bear moral responsibility in virtue of their own decision to delegate to the system; it is just that such a case would be one in which the system, *qua* proxy, could bear responsibility as well.

In a prescient paper, Bechtel (1985) considers whether there might be conditions under which we would want to attribute moral responsibility to a computational system. He argues that learning systems with the appropriate internal structures, and suitably embedded and

adaptive within the world, could respond to the demands of the world in a way that would indicate possession of the requisite rational control over their behaviour to qualify for moral responsibility. His discussion is relevant to our case because these are precisely the sorts of systems that would plausibly instantiate the behaviour in the example of a sense-2 autonomous system given above. Indeed, I have gone deeper by saying that one response to the world that we might take as indicative is an other-regarding response to the normatively forceful needs and goals of agents in the environment, which – at least in the short-term – the system would place above optimisation of its own goals.

I have already said that we would have to infer some intentionality from systems we were prepared to say acted, or even just took actions, for moral reasons. To act for reasons just is to act intentionally. Dennett (1989, 1978) is what Bechtel calls an instrumentalist about intentional states, such that ascriptions of intentional states should be made when doing so offers real explanatory and predictive power about a system's behaviour. By contrast, Bechtel is a realist about these states, such that attributions of intentional states are interpretations of the ways the system's internal configuration (its models and system architecture) prepares and enables it to have a certain kind of relationship to, or fit with, its environment (Bechtel, 1985, p. 300). He therefore treats intentional states as relational states of the system. Bechtel's central claim is that the behavioural responses of such systems would be under their control to the extent that they are how they are because of how they have adapted to their environments. In essence, this is one way of describing an autonomous system that starts to meet, at least to some degree, the ownership and guidance control requirements of reasons-responsiveness for its own actions.

One might be inclined to object at this juncture along the lines of Searle in the Chinese Room Argument that such systems would still not have any *understanding* of what they are doing – and the knowledge condition is the second condition I have said that morally responsible agents must be able to meet. We might say that the inputs which determine its outputs have no referential meaning for the system itself, and only for the people observing its behaviour (Searle 1997; Searle 1980). As one might expect, Bechtel has an answer to this, arguing that when those inputs are used by the system itself to adapt its own behaviour to the environment, then they should be thought of as having meaning for the system (Bechtel, 1985, p. 302). To return to the earlier discussion – and the argument from other-regarding function – personally costly altruistic behaviour that an adaptive system has learnt while

embedded in a wider moral system would be an expression of the system's own non-human understanding that the needs of moral patients are normatively forceful.

For those not convinced by the claim that adaptive, embedded autonomous systems of the kind I have described can be legitimately regarded as morally responsible agents, let us consider another feature of the scenario I have depicted, which may make the claim more plausible or compelling. Imagine autonomous systems that adaptively alter their plans and behaviour to help fallen pedestrians, or a range of other things – stop a carjacking, for example. We could imagine this in cases also beyond self-driving cars, such as the healthcare context or in agriculture. Their doing so might become part of their recognisable patterns of behaviour, and they may therefore be deeply integrated agents in, and which have adaptively fitted to, the wider moral system (sense-3). It is foreseeable that in such situations human agents would start to react to the systems as morally responsible agents. This picks up the approach taken by Coeckelbergh (2009) as discussed above, that we can ascribe moral agency on the basis of our experience and interactions. The point I make here is stronger though, that we could legitimately ascribe morally responsible agency when we regularly and predictably feel Strawsonian participant reactive attitudes towards these systems on the basis of their responses to us. The fallen pedestrian might reasonably feel gratitude the vehicle, and not just to the human principals whose proxy the system is. If we witnessed the system responding some peoples' needs but not to our own, it would be natural to feel resentment towards the system itself, for not interpreting *our* needs in the environment as normatively forceful. Over time, it may come facile to ask or question whether these participant reactive attitudes are *fitting*. As before, to accept this conclusion would involve some conceptual revision, because traditionally the exclusive referents of such attitudes have been human agents. I think we at least need to make *space* for the possibility of such events and developments – which I have argued are not physically impossible. And we to therefore need to acknowledge that the exclusivity dimension of the Standard View is contingently true – contingent upon the limits of current technology – but not necessarily true. Our Standard-View intuitions may one day have to give way to non-Standard notions of shared human and artificial moral responsibility. As such, my conclusion is for a soft Standard View position.

Even so, it is also true, I think, that such systems would not be fully morally responsible agents, but *marginally* morally responsible agents. Shoemaker develops an account of marginal cases of moral responsibility, for human agents having clinical depression, psychopathy,

high-functioning autism, or Alzheimer's dementia, for which our feelings tend to be that they are "eligible for some responsibility responses but not others" (Shoemaker, 2015, p. 216). Shoemaker's argument can provide a lens and orientation for thinking about the marginal moral responsibility of artificial moral agents. I have said, for example, that we might irresistibly come to have participants reactive attitudes such as gratitude or resentment towards the artificial moral agents I have described, in response to their demonstrably other-regarding behaviour towards us (or lack, thereof, if they demonstrate it towards other people but not to us). But the artificial moral agents' lack of emotional regard and empathy and phenomenal consciousness, which would no doubt be evident to all members of the moral community, might well make them ineligible for other reactions, such as disdain, or pity, or anger intended to make the system fully aware of what it has done. Likewise, we can say that these systems would bear a *degree* of responsibility as attributability, but should not be considered have the deep authorship of their actions that human agents have. And we can also say that certain forms and practices of responsibility as accountability would not be fitting. It would be senseless, for example, to use forms of punishment on an artificial moral agent that turn to some degree on the phenomenal suffering of the recipient. On the other hand, it would not be senseless to re-educate an artificial moral agent that seems to have gone awry in its other-regarding function. In short, in accommodating moral agents of this kind into our conceptual and normative frameworks, it would not be necessary to ascribe to them the full range of agential capacities that underscore the full range of our responsibility practices.

Summary of Chapter:

The Standard View, as I have construed it, has two parts. It is the view that human agents completely and exclusively bear moral responsibility for harms caused by autonomous systems. The delegation framework helps us to establish that a discernible set of humans – the human principals, both high-level and low-level principals – bear whole moral responsibility (that is, there are no gaps). This establishes my Standard View position as pluralist. In human-to-human cases of delegation, there are some cases where proxies share moral responsibility with their principals. If autonomous systems could meet the first agential condition of moral responsibility, such cases would be ones in which the second part of the Standard View – exclusively human responsibility – would no longer be true. My argument in this Chapter has been that it is not inconceivable, and indeed, it is physically possible that, with some plausibly permissible conceptual revision, there may come

a point when adaptive, embedded autonomous systems, qua artificial proxies, sometimes share moral responsibility with principals for harms they directly cause. This establishes my Standard View position as soft. To conclude, the position I defend is a soft, pluralist Standard View.

CHAPTER 6
OBJECTIONS, REPLIES, AND NORMATIVE IMPLICATIONS

In this final Chapter, I give a summary statement of my argument, and I address some key objections to it: that the delegation relation is either wrong or redundant; that the account I give is either too coarse or too fine-grained; or that it is either too demanding or it is not demanding enough. Finally, I consider the further normative implications of the argument beyond the answer it provides to the locus of moral responsibility question: the obligations it places on each kind of principal; its relation to the question of vicarious liability in law; and its role – drawing upon my discussion in the previous Chapter – within a broader project of conceptualising autonomous systems and our relations to them, and through them, to one another.

6.1 Summary of the argument

As we come, now, to the final Chapter, let me state the bare bones of my argument. This will be helpful for a consideration of natural objections that might be levelled against it, and which I have not covered, or have not covered in sufficient detail, in the preceding Chapters.

My central question is the *locus of moral responsibility* question: who is morally responsible for unforeseen harms directly caused by autonomous systems, and on what grounds? The real-world background to this question is that autonomous systems – with increasing degrees of autonomy on the three dimensions of machine autonomy I outlined in Chapter 1 – will plausibly cause harms no human agents foresaw and, in some cases, this lack of foresight will not have been negligent.

My central claim is that the relation between humans and autonomous systems is one of delegation, and that an analysis of this relation provides the most promising conceptual framework for principled ascriptions of moral responsibility to human agents, even for difficult cases, such as the reasonably unforeseen harm.

Discussions on questions of responsibility for technology often focus on harms that occur *within* accepted risk thresholds (Santoni de Sio & van den Hoven, 2018; Di Nucci, 2020, p. 196). These I would classify as *foreseen* harms. But there is also a real-world possibility of

autonomous systems directly causing *unforeseen* harm – in genuine accidents, as a result of honest mistakes, through emergent behaviour, or as an upshot of an undetected error – which undermine those risk thresholds in certain ways and therefore do not occur within them. To recall, there are two pathways by which such a harm could come about. Along the first pathway, which was denoted by Row II in Table 1, a foreseen output causes an unforeseen harm. Such cases include what I have called 'genuine accidents' and 'honest mistakes'. With genuine accidents, the foreseen output causes the unforeseen harm due to other events or features of the operating environment. Genuine accidents can undermine the risk threshold by showing it to be incomplete. Plausibly, the missile's contamination of the water supply is an example of genuine accident. Risk assessments often cannot capture the full range of hazards, particularly in such complex domains. With honest mistakes, the foreseen output causes unforeseen harm because the principals (whether low-level or high-level) have not reasoned correctly about the real-world effects of these outputs or their moral significance. Morgan's 'Scholar Bot' case is an example of an honest mistake. These are generally cases of getting the requirements wrong on the system, but presumably risk is calculated on the basis that the system's requirements have been validated and are correct.

Along the second pathway, which was denoted by Row IV in Table 1, an unforeseen output causes an unforeseen harm. Such cases include what I have called 'emergent behaviour' and 'undetected errors'. With emergent behaviour, the output is not foreseen because it is an unexpected property of the system arising from its internal (within the system) and external (within the environment) interactions. The assistive lifting robot example was a case of this kind. Such cases again may be more diverse than the scope of the risk assessment. With undetected errors, the output is not foreseen because the error (whether design or operational) which caused that output was not recognised prior to deployment. Subject to appropriate refinement of the example, we can see how the self-driving car's inability to detect objects in an anomalous weather condition could be a case of undetected design error. Their presence plausibly undermines the very assumptions of the risk threshold.

There are clear pathways to unforeseen harms, and these may not sit *within* accepted risk thresholds. So what are we to say here about ascriptions of moral responsibility?

It is worth recalling that, when it comes to the unforeseen harmful consequences caused by an autonomous system, the 'responsibility gap' argument seems to have its firmest ground.

This is the argument that no one is responsible because no one has sufficient control or knowledge of the system to be retrospectively responsible for its behaviour. Human agents will always cede *direct* control over an autonomous system in the operational domain, because we are not literally undertaking the task personally. But cases of unforeseen harm illustrate that we may sometimes also cede, or substantially weaken, *indirect* control, too. Particularly in cases of emergent system behaviour, unforeseen harms may occur with where human agents may not meet the receptiveness requirement of guidance control, where they may not have a meaningful opportunity to exercise regulative control, and the knowledge condition is suspended by fiat. Moreover, the fact that the harm was unforeseen may not be down to antecedent human negligence – the horizons of foresight do have natural limits. These are difficult cases for any position to answer.

The 'responsibility gap' problem is often presented as an argument that two necessary conditions of moral responsibility – the Aristotelian conditions – are not met by the obvious human candidates at the time, which we can call t2, of an autonomous system's causing of a harm. If meeting these conditions is a necessary condition of moral responsibility, then *prima facie*, if those conditions *were* suspended at t2, the candidates would be excused of moral responsibility. But there is a classic exception to the necessity of these conditions: one is not excused *by* the suspension of the conditions at t2 if one is responsible *for* their suspension through decisions or choices made at t1. This classic caveat is a tracing principle, and it applies to our case. The attributional responsibility of human agents for harms caused during which time they did not meet the Aristotelian condition traces back to a decision at t1, at which point they created the possibility of these suspensions at t2. This catch applies both to foreseen and unforeseen harms caused at t2.

But the argument goes deeper than this. It turns to the underlying structure of the relation between human agents and the autonomous systems to develop a specific form of the tracing principle. The underlying structure of the relation is of a relation of delegation. These very antecedent choices at t1, I argue, all serve the purpose of human handover, or delegation, to the system. After all, this is the system's 'agentive function' – it is the purpose it serves relative to human interests. The decision to delegate is, in fact, highly complex and multi-levelled. I have claimed that there are three decision points which all count as part of the antecedent decision at t1 to delegate to the system: immediate handover to the autonomous system (first

level); authorisation of the system for this purpose (second level); and the constraint and 'binding' of the system for this purpose (third level).[59]

The delegation relation, which is instantiated by the decision to delegate at t1, is an asymmetrical relation. Drawing on the wider literature, I have posited that it has three internal criteria or features – instigation, substitution, and authorisation (the last two we can count together as 'authorised substitution') – which weight responsibility with the principals who both instigate handover of decision-making function to the system and who authorise the system as a fitting proxy in the real world. From the asymmetry of this relation, I derived a principal-proxy tracing principle, such that responsibility as attributability for harms caused by proxies traces back to the principal. Even though the principal is not acting personally at t2, she has instigated and has authorised the system's doing it for her, or for someone else as their proxy. The normative consequences of its being done are therefore morally attributable to her *as if* she had acted personally. The conclusion I draw from this is a strong one, since I hold this to be true for all the principals involved in the decision to delegate at t1. That is, each principal at the three levels bears responsibility as attributability for the harms caused by autonomous systems; each principal is open to moral appraisal.

Thus, we have the first part of the answer to the *locus of moral responsibility* question, the locus (or loci) of moral responsibility as *attributability*:

> *The delegation relation (specifically, its three normatively asymmetrical internal criteria), which yields the principal-proxy tracing principle, provides the property, P, of the relation the principals S bear to all consequences of deployed autonomous systems' outputs, in virtue of which the attribution of moral responsibility to S for UHC is warranted.*

The second part of the procedure concerns the loci of responsibility as *blameworthiness*. The decision to delegate at t1, to which attributional responsibility traces back, incorporates a commitment to risk; indeed, it is a commitment to impose risk. And it is a commitment to impose a risk of a certain kind – a risk of harm from a machine over which the Aristotelian conditions may well be suspended at t2. The nature of the risk behoves including moral considerations within assessments of risk from deployment; assessments not just of the

[59] Given the temporal distance between the levels, and my claim that the decision to delegate at t1 involves the decision-makers at all three levels, this involves thinking of t1 as a very large 'time slice'.

overall balance of risk and benefit and the assessment that some risks are intolerably severe, but also assessments that include considerations of equity of the risk, and whether the people upon whom risk is imposed by the deployment of the system also receive the benefits from the deployment, and the power of the risk-bearing agents to resist the risk-imposition. For this decision to be non-culpable, it has not just to be non-negligent and to be made on the basis of risk evaluations that incorporate these moral dimensions, I argue. The decision also has to be made on the basis that the deployment of the autonomous system serves a purpose that morally justifies the imposition of risk at t2. Thus, we have our answer to the second part of the *locus of moral responsibility* question, the locus (or loci) of responsibility as *blameworthiness*. My conclusion here, too, is a morally demanding one:

> *Even if a) principal S has not been negligent, and b) the commitment to risk incorporates risk assessment that responds to moral values, if c) the decision to delegate does not serve a purpose that justifies the imposition of risk from delegation to an autonomous system, then d) principal S bears a relation to the risk-exposed in virtue of which S is blameworthy for all harms caused by the autonomous system, including UHC.*

The principal-proxy tracing principle is a principle for determining attribution. But I consider my rule for blameworthiness to be a heuristic principle, more of a rule-of-thumb, given the likelihood of a need for case-based reasoning about morally justifiable risk impositions. Even so, this principle and this rule-of-thumb principle are the mechanisms within the delegation framework that enable us reliably to establish attribution and blameworthiness in all cases. In this way, I use the delegation framework to defend a Standard View position, that moral responsibility for harms caused by autonomous systems is borne completely by human agents.

The delegation framework also helps us to think about future cases, when adaptive systems, embedded in real-world environments, become more independent from explicit human instruction (z-axis) than they are now – or start to develop in ways that challenge our working assumption that their agency falls short of what is required for moral responsibility. Human-to-human cases of delegation are also cases where responsibility for the consequences of the proxy's fulfilment of the task are morally attributable to the principal. But where the proxy intentionally deviates from her task – that is, she not *just* acting on orders at t2 – and causes harm in so doing, she too bears moral responsibility for the consequences. Were an artificial

proxy, an autonomous system, non-exempt on agential criteria from moral responsibility ascriptions, to do the same, we would be warranted in making a claim of shared principal-proxy responsibility in these cases, too. I argue that it is both conceptually coherent and physically possible, even if not currently technically possible or empirically likely, that autonomous systems develop in ways in which at least ascriptions of marginal moral responsibility are fitting, and fulfilment of some of the agential criteria would be undeniable, given some conceptual modifications. A delegation framework could still ascribe moral responsibility on warranted grounds to agents in such cases.

In characterizing the Standard View, I made two distinctions within the central claims that such a position might make. The first distinction is what I called the 'monist/pluralist' distinction – the distinction between whether one kind of human agent or many bears moral responsibility for harms caused by our technological systems. The second distinction is what I called the 'hard/soft' distinction – the distinction between whether human agents necessarily bear exclusive moral responsibility or whether this is only contingently true. My position is pluralist; I endorse shared human moral responsibility. In allowing that autonomous systems *might* develop in ways that warrant some ascription of moral responsibility for some of the harms they cause, my position is soft, too. Thus, my argument defends a soft, pluralist Standard View.

6.2 Objections and replies

There are several routes an objection to this argument might take. I look at three of these in the sub-sections below. The first is the objection that the delegation framework is essentially redundant, because we can derive satisfactory ascriptions of moral responsibility to human agents for harms caused by autonomous systems in a simpler way. The second objection concerns my individuation of the principals. The third objection considers whether the account I give is too morally demanding, both in the attribution of moral responsibility to all principals, whatever the circumstances, and in setting the bar for blameworthiness far lower than negligence.

6.2.1 Is a delegation framework really required?

The main rival to the delegation framework as a framework for answering the *locus of moral responsibility* question in a way that also yields a Standard View position is the instrumentalist framework. This framework, derivative upon the classical view in the philosophy of technology that all technology artefacts and devices are 'mere' tools, would trace attributions of responsibility back to the toolmakers and the tool-users. In some cases, applying the instrumentalist framework to determine the loci of responsibility could result in attribution to the same actual people.[60] If an understanding of the relation between humans and autonomous systems as a relation between toolmakers and tool-users can yield a satisfactory result – if it can provide a defence and explanation of our Standard View intuitions about where moral responsibility belongs – surely the delegation framework is redundant, or so this objection might go. In Chapter 4, I went some way to defend the account against this objection (see §4.2). I argued that to ascribe moral responsibility to toolmakers and tool users in cases of non-negligently unforeseen harm would not necessarily be grounded in desert. But the delegation framework, and the principal-proxy tracing principle, explains why such an ascription would be deserved.

To make this point clearer, we can consider the instrumentalist's precise grounds for answering the *locus of moral responsibility* question. Recall, this question specifically concerns the unforeseen harm. What would be the property, P, on this instrumentalist framework in virtue of which the toolmakers and tool-users, S, would be morally responsible for UHC? It could be grounded in the counterfactual dependence of the systems upon its makers and users. But to ground it in this alone would be to assimilate causal responsibility and role responsibility to retrospective moral responsibility, and it is not clear that they can be assimilated in this way. The instrumentalist might say that causal responsibility captures the 'instigation' element of the delegation relation. We could add to this that role responsibility captures the 'authorisation' element, since toolmakers, for example, vouch for their products and their professional rigour when making them. Perhaps, then, as I discussed in Chapter 4, it is simply part of the role responsibility of these people (and of tool users) to be ascribed moral responsibility for the harms their tools cause. If the direct cause of the unforeseen

[60] One difference is that I also include regulators amongst the high-level principal, in virtue of their role in binding and authorising autonomous systems as artificial proxies, but the instrumentalist framework could easily be extended to include regulators within its schema, perhaps on the grounds that regulators influence the toolmaking and tool use.

harm were a mere simple tool, like a dishwasher, the fact that its use was instigated by the tool user, and its safety was vouched for by the toolmaker, would ground the user's and the tool-maker's attributional responsibility for any harm caused by the dishwasher them. Why would it be different if the cause were a self-driving car or a surgical robot? If it would not be different, then delegation is unnecessary for satisfactory ascriptions of responsibility.

I have said that the system's agentive function is precisely to be a proxy, to be delegated to, to substitute the human actor in the operational domain. Irrespective of whether this is descriptively accurate (which I think it is), substitution could still be an unnecessary property P, of the relation S bears to UHC, for the question of moral responsibility to be answered. But at the heart of my claim, which is captured by the delegation framework and not by the instrumentalist understanding, is that the substitution element provides the final piece of the puzzle to derive the principle that normative consequences of the delegated task at t2 are attributable to me *as if* I had carried out the task personally. Instigation and authorisation alone do not provide this. The instrumentalist framework, which omits the element of substitution, cannot get us to this result.

Given that there are several plausible pathways by which autonomous systems could cause unforeseen harms, and given the moral significance of those harms, it is a compelling benefit of the delegation framework that it can show how ascriptions of moral responsibility to discernible human agents would be fair, or morally justified, in these difficult cases. It is further benefit that what makes ascriptions of moral responsibility warranted in these cases is the same as what makes ascriptions of moral responsibility warranted in the more straightforward cases of foreseen harm. That is, the delegation framework generalises and is flexible in a way that the instrumentalist framework struggles to achieve. Principals are deservedly morally responsible because that is inherent to the asymmetry of the relation. Because of their authorisation of autonomous systems as *proxies*, the principals are as morally responsible for the consequences as if they had acted personally, even though they literally did not.

The delegation framework also has a certain fluidity that the instrumentalist framework lacks. In this thesis, my working assumption has been that today's and the immediately foreseeable future's autonomous systems – self-driving vehicles, autonomous missiles, assistive robots, nursebots, and so on – have minimal agency. They are causal agents, with the power to effect

change in the world, and they are goal-directed agents, whose behaviour can be given a teleological explanation in terms of their objective functions, or reward functions, or fitness functions. But it remains an open question whether they are intentional agents, with literal beliefs or desires or intentions, or the capacity to act for reasons. The further normative advantage of the delegation framework is that it can accommodate autonomous systems with more-than-minimal agency. I have also shown that there may come a point when even those who are sceptical of attributing intentional agency now find it undeniable or irresistible to do so, on the basis of our interactions with them and should they learn to become sensitive to our needs and goals as members of the moral community, and should they demonstrate functionally equivalent other-regarding behaviour. At such a point, perhaps it is our conceptual frameworks that should expand to accommodate them, rather than leave them out for failing to be sufficiently humanlike in their properties. Should such systems come to pass, the Standard View position that not only are humans completely morally responsible for harms caused by autonomous systems (i.e. that there are no responsibility gaps) but that humans are also exclusively morally responsible (i.e. that only humans are morally responsible) would have to give way to a non-Standard View position. To reiterate, it is my acknowledgement of this possibility that renders my position a *soft*, pluralist Standard View. But given this background conceptual and physical possibility, the fact that the delegation framework can ascribe moral responsibility on the same grounds in these speculative, future cases provides further reason to endorse it.

Another objection might be that I have under-represented the influence of the autonomous system. This objection could take one of two routes. It might question the asymmetry I have taken to be inherent within the delegation relation. Is it not the case that autonomous systems might instigate tasks with us, even delegate to us? As Di Nucci says,

> *"artifacts need not only be delegates: they can sometimes take the role of delegating party ... [when] some computer system or algorithm decides which task to assign to which employee ... whether you get a certain job is itself decided, it could be argued, by the algorithm."*
>
> (Di Nucci, 2020, pp. 68-69).

To be clear, Di Nucci's point in raising this example is to highlight that we should be wary of implying a certain power structure between principals and proxies. Normatively asymmetrical it may be, but delegation is not necessarily a one-way relation. Another example

of delegation from system to human could be transition demands issued to human operators by autonomous systems in emergency situations, as self-driving cars will be designed to do. This would involve the human agent substituting the system in the fulfilment of a critical part of the overall driving task. But this still leaves open the instantiation of the authorisation criterion, and questions about the kind of agential capacity required to authorise another agent in one's name. Even so, even if we are happy to speak in terms of technological systems, specifically autonomous systems, being able to instigate the fulfilment of tasks to human agents as their authorised proxies, this does not impugn the fact that those systems themselves, if they are fulfilling their agentive function, are still *our* delegates or proxies. As such, we can deal with this possible worry, and its implications for moral responsibility, by tracing back to the human principals of the delegating autonomous system. Further, drawing on the analysis in Chapter 5, if we would be willing to attribute the requisite agential capacity to such systems, we might say that cases in which our artificial proxies delegate to us would be another case of shared moral responsibility. But I think we would not be warranted to go so far without considerable evidence of the system's capacity to act for moral reasons.

The second route could be the objection that I have failed to consider the fact, well-articulated by Latour and Verbeek, that the proxy also *transforms* the principal in a deeper, more existential sense (Verbeek, 2005; Latour, 1994; Latour, 1992). In the light of this position, proxies are not neutral representatives, but value-laden mediators in our human and social lives. Certainly, the influence of autonomous systems on human behaviour should not be under-stated. A whole empirical research area of 'human-machine interaction' is replete with fine-grained accounts of their psychological and behavioural impacts. The systems' interactions in our environments will also have a generative effect and give rise to new forms of morally relevant action (Floridi 2016, Floridi 2013). Autonomous systems will 'nudge' us into certain kinds and forms of behaviour (Burr, Christiani & Ladyman, 2018). Perhaps delegation under-estimates the extent and influence of machine agency, such as it is, in our lives. Nonetheless, the delegation framework also stands up against this kind of objection. It is one thing for the autonomous system to be influential, to initiate, or to prompt human decision-making and action; it is another for it to be responsible for the consequences of this. First, of course, as above, agential capacity would need to be established. But even if it was established, these influences and effects are accidental; there is no intent to be transformative from the systems, and these kinds of transformative systems do not authorise human replacements of themselves in the operational domain. In fact, a delegation

framework can accommodate the influence of autonomous systems on human agents and in generating new types of action. These influences could be part of the background considerations made during decision-making at the third and second level of the decision to delegate, for example. They could also be incorporated into the moral dimension of risk evaluation, since some (though not all) of these insights have implications for the autonomy of human users and risk-bearers.

The influence and power that technological systems have over us need not impugn ascriptions of moral responsibility to human agents for their consequences, even those that are unforeseen. Even so, it should be noted that my whole argument has been based on the stipulation (see §2.1) that each principal makes her decision within the complex decision to deploy, or to delegate, voluntarily. In many respects the voluntariness of 'decision to delegate' has a time-limited extension. We may not always make all these decisions uncoerced. It is sadly foreseeable that at some point we may not have the power to opt-out of the choice to use autonomous systems in our work or in our personal lives, as we travel, or to be recipients of them in public services. But if anything, this merely reinforces the urgency of our task, and the fittingness of moral demandingness now. If it is going to happen, it had better be on the basis of a sound and compelling moral justification.

6.2.2. The individuation of the principals

I have spoken of the principals in the singular, as individuals. This has primarily been for the sake of brevity, and to reduce complexity when analysing the structure of the delegation relation and of the complex, multi-levelled decision to delegate to an autonomous system. But it gives rise to further questions. It is time to be more explicit, and to explore this complexity a little, and to consider whether it bears on the success of the delegation framework for ascribing moral responsibility.

To recall, I have said (see §2.4.1), at each level of the decision to delegate, but particularly amongst the high-level principals, the principals will usually be members of *plural agents*, working within larger organisational structures. By plural agents, I mean groups of individuals at the respective decision-points of the complex decision to delegate to the system, pursuing the joint project of doing so, and with broadly shared intentions directed at the same goal. These joint projects would respectively be, at each level, starting from the third: to 'bind' the

system so that it is a safe and fitting proxy for a human in the operational domain; to approve the system for this purpose; and to hand over decision-making function in the operational domain. And I have also stated that I endorse 'methodological individualism' with respect to group agents, such that only individuals and not single group agents bear moral responsibility.

This might prompt the objection that I have not done enough to consider the moral responsibility of single group agents, such as corporations, within which the individual and plural agents sit, and which are not taken to be reducible to their individual members. In §2.4.1, I argued that it was conceptually apt to exclude single group agents from consideration on the grounds that, even though it would be legitimate to say that such agents can be reasons-responsive (and hence able to meet the control condition), any talk of single group agents understanding what they are doing (and hence able to meet the knowledge condition) would be merely elliptical for the understanding of individual humans within the single group agent. Moreover, because single group agents are not human agents, I adopted the strategy of excluding them from the Standard View. My project has therefore been to establish a framework for ascribing moral responsibility to discernible sets of individual human agents for harms caused by autonomous systems.

To this conceptual claim, I now add a claim that ascribing moral responsibility to single group agents in these cases – and certainly exclusively to single group agents – has no distinctive explanatory power. Recall the question asked in Chapter 1 (§1.2) where, given that human agency has been involved in the creation and deployment of the systems, it would be natural and fitting for anyone harmed directly by a system to ask: Who did this? Who is to blame? An answer to the effect that corporations and institutions X and Y are morally responsible would very likely elicit demands for further explanation from the addressee, such as: What decisions led to this? Who took those decisions? Why? By focusing on the locus (or loci) of individual moral responsibility, we get to the heart of the addressees' concerns and provide the more satisfying answers and explanations that they seek and deserve.

Even so, it is interesting that the delegation framework (though not the Standard View as I have characterised it) can accommodate a position that holds – contra mine – that moral responsibility can be borne by single agent groups such as corporations. First, it could allow that such irreducible group agents might be principals. The delegation framework could be

extended to include them. We might want to say, for example, that the principal at the second level of the complex decision to delegate is a whole regulatory body, such as the Medicines and Healthcare products Regulatory Agency (MHRA). Second, it could allow that single group agents might bear moral responsibility in addition to that which is borne by principals working together in plural groups. Even if, for example, some whole corporation could be held morally responsible for outcomes, that does not mean that there is no role for personal moral responsibility any more.

A second possible objection with respect to the individuation of the principals is the fact that each of these individuals and plural agents will be situated within in highly complex networks and chains of decision-making, spanning great distances and long periods of time. The harm caused by undetected error, for example, which was one of my Row IV cases, might have occurred in the manufacturing of a semi-conductor, for example, which is far removed from the decision at the second level decision point to approve a system for use in its intended context, and commit to the risk of so doing. Does this intricacy, and plurality of human agency, and this remoteness and distribution of agency mean that my characterisation of classes of principal is not only optimistic in its simplicity, but also problematic for a delegation account of moral responsibility for harms caused by autonomous systems?

This prompts me to sharpen my account. I have said, throughout the thesis, that the advantage of the delegation account is that it enables us to identify discernible sets of individuals at three decision-points. But we can draw the boundaries of these decision-points more precisely – to the principals, usually within plural group agents, who make the *final* relevant decision with their level. Let me be clearer. At the third level, these principals would be those who, at the end of the process of binding the system for delegation, confirm, often as members of a group, that this has been done sufficiently well to proceed with the decision to deploy. At the second level, the principals would be those who, after setting the rules for authorisation or after certifying a system, likewise confirm that this has been done sufficiently well to proceed with the decision to deploy. And at the first level, the principals would be those who decide to deploy in the real-world. In some contexts, these low-level principals may not be the individuals who immediately hand over decision-making function to the system, but key decision-makers such as the care home manager or the military commander who authorise the handover as a policy. Thus, by tracing responsibility as attributability back to three key decision points – binding for delegation, approving for delegation, and

delegating – our procedure can be focused. But of course this does place an obligation upon the principals to ensure that the decision-making at these three decision-points incorporates and evaluates the outcomes of the decisions of other agents distributed across the network, particularly at the second and third levels of delegation.

Even given this procedure, however, there would likely yield a substantial number of individual principals within each plural agent, and possibly even more than one plural agent at each decision-point, within the complex, multi-levelled decision to delegate to the autonomous system. Even if each individual is personally responsible *qua* principal, perhaps moral responsibility is still to highly *diffuse* – that is to say, scattered or distributed across too many agents. This diffusion would be a worry if it impedes our ability to make ascriptions of moral responsibility that satisfy their addressee. Imagine that the child's parent is told that the members of three or more sets of principal plural agents at three decision points are all morally responsible for the injury. It seems unlikely that she would be comforted by the thought that at least she now knows who to blame.

To recall, however, I have said that the principal-proxy tracing principle provides a principle for moral responsibility interrogations for harms caused by autonomous systems. I stand by the claim that each principal is attributionally responsible since they are rightly open to moral appraisal for their part in the decision to delegate to an autonomous system. In one way, this is a morally demanding conclusion; it says that each principal is open to moral appraisal for a harm that may be spatio-temporally remote from that decision. Worries about *over*-demandingness are discussed in the next sub-section. But I have also said that there is scope here for substantial case-based reasoning too, looking at the particularities of each case and the relations between the principals (see §4.2). One dimension of this more particularist reasoning would be to determine whether each principal bears attributional responsibility to the same *degree*, and to consider what would inform these judgements. This development, while still morally strict or demanding, would have two benefits. First, it would help to assuage worries that responsibility is distributed across too many principals. The addressee, for example, would likely be more satisfied if told that some of these principals bore attributional responsibility to a greater degree. Second, it would be more just, because it would help to ensure that those who played a weightier role in the decision to delegate would bear more of the weight of attributional responsibility.

6.2.3 Demandingness objections

Throughout this thesis, I have been clear that my argument is a morally demanding one. More specifically, my account of responsibility as attributability is demanding, because it holds that principals are attributionally responsible for the consequences (including the unforeseen harm) of delegating to an artificial proxy as if they had caused the harm personally, and my account of the responsibility as blameworthiness is demanding, because it holds that these principals would be blameworthy only if they were not praiseworthy for delegating to the autonomous system in the first place.

A natural objection might be that my argument is in fact too demanding. Let me first defend my account against an objection of the over-demandingness of my responsibility as attributability claim. This is the claim that the delegation relation is sufficient to ground attributions of moral responsibility to all the principals, even for unforeseen harms. My argument has been that such attributions would be deserved (and would not, as the alternative instrumentalist framework seems to require, be arbitrary). The objection here would be that the normative consequence of my responsibility as attribution claim would be intolerably demanding, and hence unfair.

To recall, the main reason why I said it would be deserved is because principals, through the decision to delegate, authorise the autonomous system to act in their place: it does so in their name. This is why the consequences of the system's doing so, such as causing a harm, are as attributable to them as if they had done it personally, even though it is clear that the machine and not they directly caused that harm. Let us recall also that responsibility as attributability means that attributionally responsible agents are called upon to answer for their decision to delegate to the system that went on to cause the harm and that they are open to moral appraisal on account of the answers that they give. The putative intolerableness of this demanding account must surely be the 'as if they had done it personally' dimension of the attribution claim. But my claim is first, that this is simply entailed by the nature of the delegation relation, and second, that surely it is not intolerably demanding to ask principals who make this morally significant decision to delegate to an artificial proxy at least to provide an answer for, to provide a justification for, that decision. But perhaps the supposed intolerableness resides in the fact that potentially a vast number of principals would be attributionally responsible in this way – and perhaps this triggers an ought-implies-can

principle to the effect that they could not even be located let alone morally appraised. This is where the sharpening of my account in the previous sub-section becomes salient: drawing the boundaries of decision-points such that it is the principals who make the final relevant decision at each of the three levels to whom responsibility as attributability traces. It would not be too practically difficult to locate these principals; moreover, a prospective duty could be imposed such that these principals are required to name themselves and, particularly at the third and second levels (but also at the first level when the decision to delegate is not *simply* made by an individual lay user of an autonomous system), to document their decision-making with respect to the binding of the system, the certification or authorisation of the system, and the handover to the system, respectively.

Let us now turn, second, to a defence of the demandingness of my responsibility as blameworthiness claim. Specifically, my claim is that, as a rule of thumb, only if they were morally praiseworthy in their decision to delegate to an autonomous system, would the principals not be morally blameworthy for the harmful consequences of doing so. This means that all cases in which, even if there was no negligence, and even if, commensurate with the nature of risk created by the deployment of autonomous systems, moral values were included in the risk assessment, if the decision was not made for good moral reasons (which I have cashed out as that it serves a purpose that morally justifies the risk imposition), the principals would deserve moral blame for even unforeseen harms.

I think that the low threshold for blameworthiness proposed is defensible, given the nature of the risk imposed and the background conditions of uncertainty. Expectations upon one to have good moral reasons for imposing risks must surely increase where those risks are morally serious. The commitment to a morally serious risk should not be a gamble, even one that has been constrained by the first two requirements of my blameworthiness heuristic, but a morally conscientious decision, made for the right reasons. And the risk in cases involving autonomous systems *are* morally serious. Not only are they risks of harm caused by a machine, and which may therefore be particularly morally and psychologically damaging to the harmed individual, but they are also risks from machines that may behave in surprising and in unforeseen ways, and over which there is weakened indirect guidance control, and possibly not even meaningful regulative control over the pathway from the output to the harm. In addition, because the system is not carrying out the task as a human would, by the decision to delegate, the principal instigates a whole new range of possible, undesirable

consequences which may not have occurred if she had not made the decision to delegate to an artificial proxy. With features such as these, demanding that the risk be imposed for morally good reasons is appropriate.

In the previous sub-section, I sharpened by account by introducing the feature of case-based reasoning. This feature is operative here, too. If, on the basis of case-based reasoning, it is determined that *some* of the principals bear a greater degree of responsibility as attributability for the harm than others, and as such a greater attributability of the commitment to risk, then those principals would commensurately bear a greater degree of responsibility as blameworthiness for the harms, too. The morally strict threshold for blameworthiness would be disproportionately borne by those fewer key individuals. Individuals with a less important authorial role, with a lesser degree of attributional responsibility, would still be blameworthy insofar as they met the conditions – but as a proportion of the overall shared blameworthiness for the harm, this would be a small and not unjust burden to bear.

Further, there is a comparative normative disadvantage of a less demanding account, which applies to both parts of my argument (i.e. both the attributability claim and the blameworthiness claim). This is the disadvantage that a less demanding account might facilitate a phenomenon known as 'agency-laundering', which has been described as follows: "obfuscating one's moral responsibility by enlisting a technology or process to take some action and letting it forestall others from demanding an account for bad outcomes" (Rubel, Castro & Pham, 2019, p. 1018). An account that (within the newly drawn boundaries of the delegation decision-points) requires that each principal answers for and is open to moral appraisal for their part in that decision to delegate prohibits such evasive tactics. And an account on which judgements of blameworthiness are made according to the moral reasons for which that decision was made and the purpose the risk-imposition serves further buttresses against the danger that nefarious intent can be obscured in apparently neutral cases.

An objection here may be raised that the demandingness of my account would seem to endorse a kind of paralysis, whereby no one could ever do anything that imposed even a small risk or chance of harm on others without a moral justification. As we have just seen, however, even if they are unlikely, these are morally serious risks. Furthermore, with widespread deployment of these systems, such that, perhaps, unforeseen harms become

more quotidian, we have the problem of systematisation. All systems of a type may cause harm in this way, so there would be many more potential risk-bearers facing these unlikely but morally serious risks.

To conclude, while my account – both the responsibility as attributability dimension and the responsibility as blameworthiness dimension – is morally demanding, it is not overly demanding, but in fact appropriately demanding.

The alternative view might be that my account is not morally demanding enough. A moral version of this objection could be that, given the risk I have described, there could be no sufficiently strong moral justification to decide to deploy – that is, to decide to delegate to – an autonomous system. Perhaps we should be even more demanding than I have maintained and say that this decision would *always* be morally impermissible and that principals would always be blameworthy for the consequences of so doing. This, I think, goes too far. After all, we have seen that some autonomous systems could positively advance human welfare in important ways – for example, by saving lives that would be otherwise lost, or by relieving humans of physically and psychologically arduous burdens. If the attendant risk-imposition is also suitably equitable, constrained, and managed, it seems strange to make it impermissible to try do as much good as one can.

A practical version of the 'not demanding enough' objection could be that any principal could come up with a narrative that their decision to delegate, all things considered, was morally praiseworthy. This practical worry may prove to be true. Indeed, it litters the marketing material of the giant technology corporations. But a distinction should be drawn between whether they do in fact meet the morally demanding threshold I have proposed (and hence are not blameworthy), and whether they can persuade people that they do. The latter is irrelevant to blame*worthiness*. Furthermore, in articulating the third and strongest part of my blameworthiness account, as the requirement that the purpose the system serves morally justifies the risk-imposition, part of the recourse such evasive principals might have to evidence for their claim is lost (see Rubel *et al.* 2019). They might, for example, try to justify the overall praiseworthiness of the decision to delegate on the basis of discrete morally relevant properties of the autonomous system, such as its impartiality or neutrality, or its energy savings, but it is less easy to evade the question of whether the very purpose served

by the system is morally compelling, and more so than alternatives to achieve that same purpose.

6.3 Normative implications

I now consider some further implications of my account, beyond its promise in providing a framework for making fair or warranted moral responsibility ascriptions for harms – whether foreseen or unforeseen – caused by autonomous systems.

6.3.1 Relations between principals and their prospective duties

Attribution concerns the relation of principals to the harm-causing autonomous system. But this has implications for the relations between the principals. Given that, on the account I advance, a harm caused by the system will be attributable to each of them (within the sharpened boundaries of the decision-points in the decision to delegate, and even if some to a lesser degree than others), each principal's openness to moral appraisal is dependent upon other principals who have participated in the complex decision to delegate to the system. Each of their decision-making is either informed by the other (as the low-level decision to hand over to the system is informed by the binding of the system and its authorisation by high-level principals) or has consequences due to the other (as low-level decision-making implements the outcomes of high-level decision-making). For delegation to the system to be successful, therefore, there should be some prospective duties on the principals in view of the shared obligations that they have not just within their plural groups but to principals at other levels of the multi-levelled decision to delegate to the system.

Particularly acute is the relation between low-level and high-level principals in respect of the fact that, while the system substitutes the human operator in the domain, the developer (as constrained by the regulator) determines how it will do so. Because of the moral relevance of the outputs of autonomous systems, some commentators have considered their characterisation as moral proxies acting on behalf of a person (Millar, 2015; Thoma, 2021). In particular, Thoma examines whether the system, such as a self-driving car, should be the moral proxy of the low-level or high-level principal, implementing outputs which low-level principals would deem morally desirable or those which the high-level principals deem correct. One aspect in which they might diverge is in differential approaches to risk, with

individuals generally, and sub-optimally, more risk averse in individual problem scenarios (Thoma, 2021). Whether or not such things as customisable ethics settings on autonomous systems for low-level principals would be desirable, this will naturally not extend to the full range of the system's functionality. But the question does reveal the importance of the sensitivity of high-level principals to the requirements of low-level principals in the decision-making process about the development of the system. As we saw in Chapter 3 (see §3.3.1.1), the degree to which the system is receptive to low-level principals' reasoning will be derivative upon their inclusion or representation in the decision-making of designers and engineers. A plausible prospective duty on the principals at the third decision point would be to ensure that the system is only deemed satisfactorily 'bound' for deployment as an operational proxy of the low-level principal if there has been adequate attention to the needs and interests of the low-level principal in the binding process.

It is clear also that another prospective duty on the high-level principals will be to ensure that the low-levels principals have accurate, accessible information about the system they delegate to, full awareness of its limitations and intended context, and understanding of their own obligations for the safe and ethical deployment of the autonomous system. By the same token, the low-level principals would have a prospective duty to delegate to the system within those agreed and intended parameters. Such considerations are of course not new. Many of the prospective duties or role responsibilities on principals that a delegation framework would give rise to – to be honest with one another, to warrant each others' trust, to be rigorous and non-negligent, to communicate accurately – are not particularly unique to the delegation framework. They would underscore all 'responsible innovation' (Owen *et al.* 2013) and professional codes of ethics. But the framing of the relation in terms of delegation emphasises their importance to the principals, because of the responsibility as attributability that I have argued the principals *share* under a delegation framework.

6.3.2 Relations between principals and the risk-exposed

Blameworthiness concerns the relation of the principals to the persons harmed, and more generally to those who have the risk of harm imposed upon them from the deployment of an autonomous system. For delegation to be non-culpable on my account, the acceptable risk thresholds determined by the principals (the principals at the second decision-point are the most implicated here) should include appraisals of moral or ethical risk. This, I think,

places a prospective duty on the relevant principals to go beyond purely technical risk assessment methods that focus exclusively on the probabilities and severities of hazards and unwanted events, and to include considerations of the agency of those exposed to risk and the benefit they themselves derive from the imposition of risk (Hansson, 2018). Indeed, this requirement also has implications for the relations between principals discussed in the previous sub-section. Given that for each principal their possible blameworthiness for harms caused by autonomous systems depends in part on whether moral values have been included in the risk assessment – or, to use Hansson's phrase whether an 'ethical risk analysis' has been performed – it seems incumbent upon all of them to ensure that this does occur.

I have said that, in addition to not being negligent, the principals would need to show that they had incorporated these sorts of moral considerations within their evaluation of risk in order not to be blameworthy for harms caused. But there was a further requirement, such that these two previous requirements alone are not sufficient *not* to be culpable. This further requirement is that the purpose served by the deployment of the autonomous system morally justifies the imposition of the risk. For there "may be some machine learning systems that should not be deployed in the first place, no matter how much we can optimize them" (Zimmermann, Di Rosa & Kim, 2020). A further prospective duty on principals, therefore, would be to include wider societal voices in particularist reasoning about why the system is being deployed in the first place, and include the perspectives and requirements of those upon whom the risk of deployment would be imposed. Amongst scholars of algorithmic justice in particular, this wider involvement of citizens, viewing the question as "a collective problem for *all of us* rather than a technical problem *just for them*" has been called the "second wave" of thinking about responsibility (Zimmermann, Di Rosa & Kim, 2020). In terms of our delegation framework, the prospective duty can be made more specific. High-level principals at the second level of the decision to delegate – that is, the regulators – should establish robust and effective mechanisms for this consideration of the moral justification for the deployment of autonomous systems to take place. Indeed, as before, given that each principal has a stake in this, it is plausible that a prospective duty on the principals at the other two levels of the decision to delegate should ensure that this happens, or that it has happened, prior to making the decision to delegate, or to deploy – which is also, as I have said, the decision to commit to the risk of so doing.

As in my discussion in the previous sub-section, this in one sense is not new. Consultation papers for wider review amongst the public, citizen's juries, and participatory engagement are already established procedures of good practice in technological development (Owen *et al.*, 2013). But what the delegation provides is a clearer focus for these activities, and a framing question: would the decision to deploy, which is also the decision to delegate, which is also the decision to commit to the risk of this, be a morally praiseworthy decision, all things considered?

6.3.3 Moral justification for vicarious liability

I have not engaged, in this thesis, in questions of legal responsibility, even though there is an isomorphic debate in legal scholarship about the loci of legal liability for harms caused by AI-based and autonomous systems. Even so, as raised in Chapter 2 (§2.1), the delegation framework for ascribing moral responsibility to principals has parallels with a mechanism known as 'vicarious liability' in law: "which create[s] responsibility for one person, the 'principal', for actions undertaken by another person, the 'agent'" (Turner, 2019, pp. 98-99). Vicarious liability is an "interesting subspecies" of strict liability, which is liability "for which the contributory fault condition is weakened or absent", where 'contributory fault' means that the liable agent's negligence was causally contributory to the harm (Feinberg, 1970, pp. 222-223). Vicarious liability is most prominent in tort law, but it is also, far less often, operative in criminal law.

Vicarious liability in tort law requires that there is a relationship between the liable agent and the tortfeasor (the person who has committed the offence that harms or injures another party) that is sufficient to trigger the doctrine. Typically, vicarious liability occurs in employment contexts, whereby employers are vicariously liable for the torts of their employees, so long as the offences in question are connected to their employment. But it applies to other kinds of relationship, too, and particularly where the liable agent has some form of power or control over the tortfeasor, such as the ability to tell her what to do, and the tortfeasor's work is integrated into the liable agent's organisation (Burton *et al.,* 2019, p. 10)

To recall, liability – which is about making restitution, such as paying compensatory damages to the victim – is an aspect of responsibility as accountability. My concern in this thesis,

however, has been with responsibility as attributability, and judgements of blameworthiness that might be made as an upshot of the moral appraisal of attributionally responsible human principals. However, in taking the central relation between human agents and autonomous systems to be a relation of delegation – a relation between a principal and a proxy – the account I have advanced could support positions in legal scholarship that identify the relationship between certain human agents and autonomous systems as one that could, with some modifications, trigger the doctrine of vicarious liability and which take the vicarious liability model to be well-suited to "the unique functions of AI which differentiate it from other man-made entities" (Turner, 2019, p. 90 & p. 101).[61] More precisely, my account could support arguments that hold that a form of vicarious liability for autonomous systems is a plausible and promising way to close the legal liability gap – which is the gap that arises when a harm is caused by an autonomous system but the loss falls on the victim of the harm (see, for example, Burton *et al.* 2019, pp. 9-12). This is an interesting normative implication.

It is worth recalling that in general what makes an ascription of responsibility 'strict' is that the ascription is made in the absence of fault – where in this context that means absence of negligence in the liable party.[62] The legal philosopher Hart clearly thinks that vicarious liability is also independent of moral blameworthiness, and is, indeed, an exemplar of the dislocation that may occur between legal responsibility and moral responsibility (Hart, 1968, p. 223). Since I pursue the further question of blameworthiness even in the absence of negligence, my account may also advance a moral justification for vicarious liability – at least in these cases. This has relevance for a debate within the philosophy of law. My argument is that the principal-proxy tracing principle traces moral responsibility as attributability back to the principals in virtue of their decision to delegate to the system, and that these principals would also be blameworthy in many cases that are not strictly negligent on account of their committing to a risk-imposition without having a good moral reason to do so. This is closer to Honoré's justification of the liability, contra Hart, which provides the moral basis of the liability in the proposition that a person who seeks gain must accept the cost of the associated risks (Honoré, 1988). Though I do not consider the connection in more detail than this here, it would be interesting to pursue further whether my emphasis on the moral justification for

[61] One such modification would be needed to deal with the fact that the tortfeasor in vicarious liability cases has to be a legal person, and autonomous systems are not presently granted legal personhood, and full legal personhood for autonomous systems is problematic because, for example, it would mean that the system would have its own legal rights (Burton *et al.* 2019, p. 12)

[62] Here I am using 'ascription' as a catch-all term to include responsibility attributions as well as ascriptions of accountability or liability.

ascriptions of moral responsibility to principals could offer relevant insights to discussions in the philosophy of law.

6.3.4 Conceptual implications

A consistent thread running through my argument has been that autonomous systems force us to confront the normative adequacy of our concepts. My discussion has incorporated two main conceptual points. First, the adequacy of our conceptual frameworks for ascribing moral responsibility for harms caused by autonomous systems. Second, the adequacy of our concept of moral agency, and how it might (or one day should) extend to include advanced and physically embodied AI-based and ML-based systems. In cases such as these, a question is whether the phenomena before us requires us to revise or reengineer some of our concepts. To recall Johnson & Verdicchio's remark in the Introduction, we are both creating the systems and asking ourselves what they are. In the process of so doing, inevitably we run up against the question of whether our ways of conceptualising the world are fitting for the phenomena. Blackburn introduces the concept of 'conceptual engineering' by explicit analogy with physical engineering: *"just as the engineer studies the structure of material things, so the philosopher studies the structures of thought"* (1999, p. 2). Eklund describes concepts as *"the tools we have come to use to understand the world"* (2015, p. 377). For both Blackburn and Eklund, conceptual engineering is fundamental to the enterprise of philosophy. Concepts constitute our structures of thought, and these structures are the frame through which we understand the world. If they fail to help us do that, or our concepts seem incapable of helping us to understand or regulate a new phenomenon in our world, then we should engineer our concepts in response, just as engineers build bridges to deal with rivers.

I have made two main claims about autonomous system which are of relevance to the conceptual revision or engineering project. First, that, no matter whether autonomous systems have intentional agency, they can still be regarded – indeed, it would be preferable to regard them – as delegates or proxies. This approach has been pioneered by a handful of scholars in recent years, but in general this way of conceptualising autonomous systems marks a shift from the dominant instrumentalist framing, and, I think, an amelioration. My second claim is that we should be prepared to admit the conceptual legitimacy of autonomous systems as moral agents. That is, I have argued that should the systems display regular patterns of behaviour that truly instantiate the form and function of moral agency

within a wider moral system, we should not rule them out from the set of moral agents on the grounds, say, of lacking consciousness or emotion.

But in terms of the frameworks for we have for ascribing moral responsibility, my argument has been that in fact we do not need a radical conceptual overhaul. Rather, we can recourse to a relation between principal and proxy that has long been recognised but, for the most part, neglected in the philosophy of action and responsibility. It roots go back at least as far as Hobbes' *Leviathan*, and the transfer of power when one authorises another to act as one's proxy, as an artificial person (Hobbes, 1651). That is not to say that my appeal to the principal-proxy relation has not involved some conceptual revision. The first is that autonomous systems are allowed into this frame. Indeed, part of the delegation relation's promise is in how easily it can accommodate them, as bound agents. The second is that I have attempted a general synthesis of the key features of the delegation relation from accounts in the wider literature of this form of shared or extended agency to identify three internal criteria of the relation: instigation; authorisation; and substitution. But, despite these revisions and development, overall, the delegation framework is not so much an exercise of conceptual engineering, perhaps, as of urgently needed conceptual reframing.

Summary of Chapter:

In this Chapter, I have considered some key objections to my account, which have prompted me to revise and sharpen aspects of it. I have set tighter boundaries on the decision-points at the three levels of the complex, multi-levelled decision to delegate to an autonomous system, and I have admitted that, while responsibility as attributability traces to each principal, principals may bear responsibility as attributability to different degrees. In addition, this Chapter has considered some of the wider implications of my argument – the prospective duties on principals if delegation is to be successful, its connection to an analogous phenomenon (vicarious liability) within the law, and the degree to which it marks a revision of our traditional conceptual frameworks for thinking about technology. Each of these would merit exploration in future work.

CONCLUSION

Delegation powerfully describes human relations to autonomous systems. It is integral to the very purpose of autonomous systems that they carry out tasks *for* human agents, and that they do so when we explicitly transfer decision-making function *to* them in the real world. This relation is morally significant, not just because of the critical nature of the tasks that are delegated to them, but because the act of delegation itself opens up the risk of a new kind of harm, caused by a machine over which, at the point of causing harm, humans do not exercise *direct* or may not even exercise meaningfully *indirect* operational and rational control.

Refocusing on autonomous systems as delegates or proxies, as opposed to thinking of them as 'mere' tools or mere extensions to automation, provides a principled mechanism for keeping moral responsibility where it seems to belong, with human agents, notwithstanding any objections about possible 'responsibility gaps'. Against a backdrop in which autonomous systems look set, over the coming decades, to be an increasing presence in our lives, it usefully provides grounds for ascribing moral responsibility in *all* possible cases involving an autonomous system.

By thinking about the reasonably unforeseen harmful consequence, and not just the harm that traces back to an obvious antecedent human fault, we can consider the subtle question of who would bear responsibility when all human agents have been rigorous and conscientious and trying to avoid harm. As Tiles & Oberdiek put it,

> *"A self-respecting scientist or engineer, in short, does not ask, 'What can I get away with without violating the law or leaving myself open to a law suit?', but 'What precautions must I take to avoid moral culpability for my conduct?' The former looks to a judge sitting on a bench; the latter to one's conscience as judge."*
>
> (Tiles & Oberdiek, 1995, p. 181)

Through the delegation relation, we can answer the conscientious actor's question. We can trace back responsibility as attributability to all human principals on warranted grounds, as rightly open to moral appraisal for their participation in the decision to delegate to an autonomous system. It is striking that, rather than give rise to worries about the weakening of responsibility, the delegation relation actually makes human moral responsibility stronger. We are responsible for more consequences when we delegate, and we are responsible for

them as if we had done it personally. In addition, I have argued, the threshold for blameworthiness in these cases is lower than the absence of negligence. Given the nature and the imposition of risk, principals would ultimately be blameworthy for a harm caused by an autonomous system if the decision to delegate to it did not serve a morally compelling purpose, which justified the imposition of the risk, all things considered. I had not expected, as I set out on this inquiry, that it would yield quite such bold conclusions.

There are several avenues in the thesis that prompt consideration for future work: how the delegation would apply in the case of advisory and not just replacement systems; how it would apply when the decision to delegate is less voluntary than it is in the early days of developing and deploying the technology; and a closer analysis of the relation between delegation and risk. Delegation is strangely underexplored as an action-theoretic concept in the wider philosophy of shared and extended agency, despite the widespread instantiation of delegation relations in society. I hope that this thesis has made some contribution to filling that gap, and that philosophers start to take more of an interest in this fascinating inter-agential relation.

I have argued that, at present, the delegation framework enables us to uphold our Standard View intuitions that moral responsibility for harms directly caused by autonomous systems rests both completely and exclusively with human agents. Specifically, I have argued that it rests with a plurality of human agents, and as such I have advanced a pluralist Standard View position. But, just as in human-to-human cases of delegation, where moral responsibility is sometimes shared between principals and their proxies, there may come a time when moral responsibility is also shared between human principals and artificial proxies. Though the autonomous systems of today and in the foreseeable near future fall short of the agential capacity that would be necessary for them to share moral responsibility, I have argued that it is neither conceptually incoherent nor physically impossible that they one day might meet this condition. As such, I have also advanced a soft Standard View position. Should such systems come to pass, the Standard View as I have characterised it, with its claim of exclusively human moral responsibility, would no longer be upheld. But the delegation framework would still apply.

The delegation framework is remarkably flexible. Not only does it extend to both foreseen and unforeseen harms, both within and outside of accepted risk thresholds, but it also

extends to possible future cases in which the autonomous systems may start to have richer and more sophisticated agential capacities. If autonomous systems do become adaptively enmeshed in our everyday lives and living environments, and start at some point to approach the minimal conditions for moral agency, the framework stands ready to accommodate them, with no loss of coverage of human moral responsibility.

Bibliography

Allen, C., Varner, G. & Zinser, J., 2000. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3), pp. 251-261

Allen, C., Smit, I. & Wallach, W., 2005. Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3), pp. 149-155

Allen, C. & Wallach, W., 2012. Moral machines: contradiction in terms or abdication of human responsibility. In Lin P., Abney K. & Bekey G. (Eds.), *Robot ethics: The ethical and social implications of robotics*, pp. 55-68. MIT Press

Alvarez, M. & Hyman, J., 1998. Agents and their actions. *Philosophy*, 73(284), pp.219-245

Alvarez, M., 2013. Agency and two-way powers. *Proceedings of the Aristotelian Society*, vol. CXIII, no. 1, pp. 101-121

Anderson, M., Anderson, S.L. & Armen, C., 2004. Towards machine ethics. In *AAAI-04 Workshop on Agent Organizations: Theory and Practice*

Anderson, M., Anderson, S.L. & Armen, C., 2005, November. Towards machine ethics: implementing two action-based ethical theories. In *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*, pp. 1-7

Anderson, M. & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4), pp. 15-26

Anderson, M. & Anderson, S.L. (Eds), 2011. *Machine ethics*. Cambridge University Press

Anscombe, G.E.M., 1957. Intention. *Proceedings of the Aristotelian Society*, 57 (1), pp. 321–332

Arpaly, N., 2003. *Unprincipled virtue: an inquiry into moral agency*. Oxford University Press

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, pp 82-115

Aristotle, 2002. *Nicomachean ethics*. Oxford University Press

Asimov, I., 1964. *The rest of the robots*. Doubleday

Axelrod R., 1984. *The evolution of cooperation*. Basic books

Audi, R., 2006. *Practical reasoning and ethical decision*. Routledge

Bechtel, W. 1985. Attributing responsibility to computer systems. *Metaphilosophy,* 16(4), pp. 296-306

Behdadi, D. & Munthe, C., 2020. A normative approach to artificial moral agency. *Minds and Machines*, 30, pp. 195-218

Bentham, J., 1789. An introduction to the principles of morals and legislation. In Burns, J.H. & Hart, H.L.A, 1970, *The collected works of Jeremy Bentham: an introduction to the principles of morals and legislation.* Oxford University Press

Berlin, I., 1969. Two concepts of liberty. In Berlin, I. *Four essays on liberty*. Oxford University Press (New edition 2002)

Blackburn, S., 1999. *Think: A compelling introduction to philosophy.* Oxford University Press

Boden, M.A., 1996. *The philosophy of artificial life.* Oxford University Press

Boden, M., Bryson, J. Caldwell, D. Dautenhahn, K. Edwards, L. Kember, S. Newman, P. Parry, V. Pegman, G. Rodden, T. and Sorrell, T., 2017. Principles of robotics: regulating robots in the real world. *Connection Science*, 29(2), pp. 124-129

Boddington, P., 2017. *Towards a code of ethics for artificial intelligence.* Springer

Borenstein, J., Howard, A. and Wagner, A.R., 2017. Pediatric robotics and ethics: The robot is ready to see you now, but should it be trusted?. In Lin, P., Jenkins R. & Abney, K. (Eds), *Robot ethics 2.0: from autonomous cars to artificial intelligence*, pp. 127-141. Oxford University Press

Bostrom, N., 2014. *Superintelligence: paths, dangers, strategies*. Oxford University Press

Bovens, M.A.P., Ford, W. & Bovens, M., 1998. *The quest for responsibility: accountability and citizenship in complex organisations*. Cambridge University Press.

Bratman, M., 1987. *Intention, plans, and practical reason*. Harvard University Press

Bratman, M., 1993. Shared intention. *Ethics*, 104 (1), pp. 97-113

Bratman, M., 2013. *Shared agency: a planning theory of acting together.* Oxford University Press

Bringsjord, S., Arkoudas, K. & Bello, P., 2006. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4), pp. 38-44

Brink, D.O. & Nelkin, D.K., 2013. Fairness and the architecture of responsibility. *Oxford studies in agency and responsibility*, 1, pp. 284-313.

Brown, A., 2009. *Personal responsibility: why it matters*. Continuum Books

Brundage, M., 2014. Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3), pp. 355-372

Bryson, J., 2010. Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: key social, psychological, ethical and design issues*, pp. 63-74. John Benjamins Publishing Company

Buchanan, B.G. & Shortliffe, E.H. (Eds), 1984. *Rule-based expert systems: the MYCIN experiments of the Stanford heuristic programming project.* Addison Wesley

Bunge, M., 1977. Towards a technoethics. *The Monist*, 60(1), pp. 96-107

Burr, C., Cristianini, N. & Ladyman, J., 2018. An analysis of the interaction between intelligent software agents and human users. *Minds and Machines*, 28(4), pp. 735-774

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., and Porter, Z., 2020. Mind the gaps: assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence*, 279, 103201

Capes, J.A., 2019. Strict moral liability. *Social Philosophy and Policy*, 36(1), pp.52-71

Carpenter, J., 2016. *Culture and human-robot interaction in militarized spaces: A war story*. Routledge

Carter, M., 2007. *Minds and computers: an introduction to the philosophy of artificial intelligence.* Edinburgh University Press

Cave, S., Nyrup, R., Vold, K. & Weller, A., 2018. Motivations and risks of machine ethics. *Proceedings of the IEEE*, 107(3), pp. 562-574

Chalmers, D.J., 2009. The singularity: A philosophical analysis. In Schneider, S., (Ed), *Science Fiction and Philosophy*, pp. 171-228

Charette, R., 2018. Michigan's MiDAS unemployment system: algorithm alchemy created lead, not gold. *IEEE Spectrum*, 24 Jan 2018. Available at: https://spectrum.ieee.org/riskfactor/computing/software/michigans-midas-unemployment-system-algorithm-alchemy-that-created-lead-not-gold [Accessed 29 June 2021]

Clark, S.R.L., 1983. Animal rights and human morality. *Environmental Ethics*, 5(2), pp. 185-188

Clarke, R., McKenna, M. & Smith, A.M. (Eds), 2015. *The nature of moral responsibility: new essays.* Oxford University Press

Clement, G., 2013. Animals and moral agency: the recent debate and its implications. *Journal of Animal Ethics*, 3(1), pp. 1-14

Coeckelbergh, M., 2009. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society*, 24(2), pp. 181-189

Coeckelbergh, M., 2010. Robot rights? towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), pp. 209-221

Coeckelbergh, M., 2016. Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence*, 30(8), pp. 748-757

Coeckelbergh, M., 2020a. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, 26(4), pp. 2051-2068

Coeckelbergh, M., 2020b. *AI Ethics*. MIT Press

Coleman, K.G., 2001. Android arete: Toward a virtue ethic for computational agents. *Ethics and Information Technology*, 3(4), pp. 247-265

Contissa, G., Lagioia, F. and Sartor, G., 2017. The ethical knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, *25*(3), pp. 365-378

Cooper, D.E., 1968. Collective responsibility. *Philosophy*, 43(165), pp. 258-268

Copeland, J., 1993. *Artificial intelligence: a philosophical introduction*. John Wiley & Sons

Copp, D., 1979. Collective actions and secondary actions. *American Philosophical Quarterly*, 16(3), pp. 177-186

Copp, D., 1980. Hobbes on artificial persons and collective actions. *The Philosophical Review*, 89 (4), pp. 579-606

Copp, D., 1995. *Morality, normativity, and society*. Oxford University Press

Copp, D., 2008. Darwinian skepticism about moral realism. *Philosophical Issues*, 18, pp. 186-206

Corlett, J.A., 2001. Collective moral responsibility. *Journal of Social Philosophy*, 32(4), pp. 573-584

Cummings, M.L. and Britton, D., 2020. Regulating safety-critical autonomous systems: past, present, and future perspectives. In Pak, R., de Visser, E.J., Ericka Rovira, E. (Eds), *Living with robots: emerging issues on the psychological and social implications of robotics*, pp. 119-140. Academic Press

Crane, T., 2016. *The mechanical mind: a philosophical introduction to minds, machines and mental representation, third edition*. Routledge

Danaher, J., 2016. Robots, law and the retribution gap. *Ethics and Information Technology*, 18(4), pp. 299-309

Danielson, P., 1992. *Artificial morality: virtuous robots for virtual games*. Routledge

Davidson, D., 1963. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23), pp. 685-700

Davidson, D., 2001. *Essays on actions and events*. Oxford University Press (reprinted 2011)

Dennett, D.C., 1978. *Brainstorms*. Bradford

Dennett, D.C., 1984. *Elbow room: the varieties of free will worth wanting*. MIT press

Dennett, D.C., 1989. *The intentional stance*. MIT press

Dennis, L. and Fisher, M., 2018. Practical challenges in explicit ethical machine reasoning. Available at https://arxiv.org/abs/1801.01422 [Accessed 25 June 2021]

Di Nucci, E. & Santoni de Sio, F., 2014. Who's afraid of robots? fear of automation and the ideal of direct control. *Roboethics in Film*, RoboLaw Series (forthcoming). Pisa University Press

Di Nucci, E. & Santoni de Sio, F., 2016. *Drones and responsibility: legal, philosophical and socio-technical perspectives on remotely controlled weapons*. Routledge

Di Nucci, E., 2020. *The control paradox: from AI to populism*. Rowman & Littlefield Publishers

Driver, J., 2008. Attributions of causation and moral responsibility. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 2. The cognitive science of morality: intuition and diversity*, p. 423–439. MIT Press

Duff, R.A., 2009. Strict responsibility, moral and criminal. *The Journal of Value Inquiry*, 43(3), pp.295-313

Dworkin, G. 1988. *The theory and practice of autonomy*. Cambridge University Press

Eklund, M., 2015. Intuitions, conceptual engineering, and conceptual fixed points. In *The Palgrave handbook of philosophical methods,* pp. 363-385. Palgrave Macmillan

Etzioni, A., & Etzioni, O., 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21(4), pp. 403-418

European Commission, 2019. *Ethics guidelines for trustworthy AI*. Available at https://digital-strategy.ec.europa.eu/en/library/draft-ethics-guidelines-trustworthy-ai [Accessed 25 June 2021]

Feinberg, J., 1970. *Doing & deserving; essays in the theory of responsibility*. Princeton University Press

Feinberg, J., 1989. *The moral limits of the criminal law: volume 3: harm to self*. Oxford University Press

Figdor, C., 2018. *Pieces of mind: The proper domain of psychological predicates*. Oxford University Press.

Fischer, J.M., 1994. *The metaphysics of free will* (*volume 1)*. Blackwell

Fischer, J.M. & Ravizza, M. (Eds), 1993. *Perspectives on moral responsibility*. Cornell University Press

Fischer, J.M. & Ravizza, M., 1998. *Responsibility and control: a theory of moral responsibility*. Cambridge University Press

Fischer, J.M. and Tognazzini, N.A., 2009. The truth about tracing. *Noûs*, *43*(3), pp. 531-556.

Fisher, M., Mascardi, V., Rozier, K.Y., Schlingloff, B., Winikoff, M. & Yorke-Smoth, N., 2021. Towards a framework for certification of reliable autonomous systems. *Autonomous Agents and Multi-Agent Systems* 35 **(**8). Available at https://doi.org/10.1007/s10458-020-09487-2 [Accessed 25 June 2021]

Fischhoff, B., Kadvany, J. & Kadvany, J.D., 2011. *Risk: A very short introduction*. Oxford University Press

Fossa, F., 2018. Artificial moral agents: moral mentors or sensible tools?. *Ethics and Information Technology*, 20(2), pp. 115-126

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication 2020-1*. Available at https://cyber.harvard.edu/publication/2020/principled-ai [Accessed 25 June 2021]

Floridi, L. & Sanders, J.W., (2004). On the morality of artificial agents. *Minds and Machines*. 14(3), pp. 349-379

Floridi L., 2013. Distributed morality in an information society. *Science and engineering ethics* 19 (3) pp. 727-743

Floridi, L., 2016. Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 20160112.

Floridi, L., 2019. What the Near Future of Artificial Intelligence Could Be. *Philosophy & Technology* 32, pp. 1-15

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5, pp. 1-7

Fossa, F., 2018. Artificial moral agents: moral mentors or sensible tools?. *Ethics and Information Technology*, 20(2), pp. 115-126

Frankfurt, H. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23), pp. 829-839

Frankfurt, H., 1971. The principle of alternative possibilities. *Journal of Philosophy* 28, pp. 339-345

Frankfurt, H., 1978. The problem of action. *American Philosophical Quarterly*, 15, pp. 157–162.

Frankfurt, H., 1988. *The importance of what we care about*. Cambridge University Press

Frankish, K. & Ramsey, W.M. (Eds), 2014. *The Cambridge handbook of artificial intelligence*. Cambridge University Press.

French, P.A., 1979. The corporation as a moral person. *American Philosophical Quarterly*, 16(3), pp. 207-215

French, P.A., 1984. *Collective and corporate responsibility*. Columbia University Press.

Gilabert, P. and Lawford-Smith, H., 2012. Political feasibility: A conceptual exploration. *Political Studies*, 60(4), pp.809-825

Gilbert M., 2013. *Joint commitment: how we make the social world*. Oxford University Press

Gips, J., 1995. Towards the ethical robot. In Ford K., Glymour C. & Hayes P. (Eds.), *Android epistemology,* pp. 243-252. MIT Press

Giubilini, A., & Levy, N., 2018. What in the world is collective responsibility?. *Dialectica*, 72(2), pp. 191-217.

Glover, J., 1970. *Responsibility*. Routledge & Kegan Paul

Goetze, T.S., 2021. Moral Entanglement: Taking Responsibility and Vicarious Responsibility. *The Monist*, *104*(2), pp. 210-223

Goldman, A.I., 1970. *Theory of human action*. Princeton University Press

Goodall, N.J., 2014. Ethical decision making during automated vehicle crashes. *Transportation Research Record*, 2424(1), pp. 58-65

Grau, C., 2006. There is no" I" in" robot": robots and utilitarianism. *IEEE Intelligent Systems*, 21(4), pp. 52-55

Gunkel, D.J., 2012. *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press

Gunkel, D.J., 2017. Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology* 22, pp. 307-320.

Habli, I., Lawton, T. & Porter, Z., 2020. Artificial intelligence in health care: accountability and safety. *Bulletin of the World Health Organization*, 98(4), pp. 251-256

Hacker, P.M.S., 2007. *Human nature: The categorial framework*. John Wiley & Sons

Hacker, P.M.S., 2013. The intellectual power: a study of *human nature*. John Wiley & Sons

Hagendorff, T., 2020. The ethics of AI ethics: an evaluation of guidelines. *Minds and Machines*, 30(1), pp.99-120

Hakli, R. & Mäkelä, P., 2019. Moral responsibility of robots and hybrid agents. *The Monist*, 102(2), pp. 259-275

Hall, W. & Pesenti, J, 2017. *Growing the artificial intelligence industry in the UK*. Department for Business, Energy & Industrial Strategy strategy paper. Available at https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk [Accessed 25 June 2021]

Hansson, S.O., 2003. Ethical criteria of risk acceptance. *Erkenntnis*, 59(3), pp. 291-309

Hansson, S.O., 2005. *Decision theory: a brief introduction*. Department of Philosophy and the History of technology, Stockholm. Available at https://www.researchgate.net/publication/210642121_Decision_Theory_A_Brief_Introduction [Accessed 25 June 2021]

Hansson, S.O., 2018. How to perform an ethical risk analysis (eRA). *Risk Analysis*, 38(9), pp. 1820-1829

Harsanyi, J.C., 1977. On the rationale of the bayesian approach: comments on Professor Watkins's paper. In Butts, R. & Hintikka, J. (Eds), *Foundational problems in the special sciences*, pp. 381-392. Springer

Hart, H. L. A., 1961. *The concept of law*. Oxford University Press (reprinted 2012)

Hart, H.L.A., 1968. *Punishment and responsibility: Essays in the philosophy of law*. Oxford University Press (reprinted 2008)

Hart, H.L.A & Honoré, T., 1985. *Causation in the law, second edition*. Oxford University Press (reprinted 2002)

Hatherley, J.J., 2020. Limits of trust in medical AI. *Journal of Medical Ethics*. pp. 478-481

Haugeland, J., 1989. *Artificial intelligence: the very idea*. MIT Press

Haugeland, J., 1998. *Having thought*. Harvard University Press

Heersmink, R., 2017. Distributed cognition and distributed morality: agency, artifacts and systems. *Science and Engineering Ethics*, 23(2), pp. 431-448

Hevelke, A. & Nida-Rümelin, J., 2015. Responsibility for crashes of autonomous vehicles: an ethical analysis. *Science and Engineering Ethics*, 21(3), pp. 619-630

Himma, K.E., 2009. Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*(1), pp. 19-29

Himmelreich, J., 2019. Responsibility for killer robots. *Ethical theory and moral practice 22*, pp. 731–747

HM Government. 2021. UK National AI Strategy. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf [Accessed 29 January 2022]

Hobbes, T., 1651. *Leviathan*. Penguin Random House (2017 edition)

Holland, J.H., 1992. Genetic algorithms. *Scientific American*, 267(1), pp. 66-73

Honarvar, A.R. and Ghasem-Aghaee, N., 2009. Casuist BDI-agent: a new extended BDI architecture with the capability of ethical reasoning. In *International conference on artificial intelligence and computational intelligence (November 2009)*, pp. 86-95

Honoré, T., 1988. Responsibility and luck: the moral basis of strict liability. In Honoré, T., 1999, *Responsibility and Fault*. Hart Publishing

Hume, D., 1751. *An enquiry concerning the principles of morals,* Schneewind, J.B. (Ed.), 1983. Hackett Publishing Company

Hyman, J., 2015. *Action, knowledge & will.* Oxford University Press

Jobin, A., Ienca, M., & Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399

Johnson, C.W., 2006a. What are emergent properties and how do they affect the engineering of complex systems? *Reliability Engineering and System Safety*, 91(12) pp. 1475-1481

Johnson, D.G., 2006b. Computer systems: moral entities but not moral agents. *Ethics and Information Technology*, 8(4), pp. 195-204

Johnson, D.G. and Miller, K.W., 2008. Un-making artificial moral agents. *Ethics and Information Technology*, *10*(2-3), pp. 123-133

Johnson, D.G. and Verdicchio, M., 2018. Why robots should not be treated like animals. *Ethics and Information Technology*, 20(4), pp. 291-301

Johnson, D. G. & Verdicchio, M., 2019. AI, agency and responsibility: the VW fraud case and beyond. *AI & Society*, 34(3), pp. 639-647

Joyce, R., 2001. *The myth of morality*. Cambridge University Press

Joyce, R., 2007. *The evolution of morality*. MIT press

Kamm, F.M., 2008. *Intricate ethics: rights, responsibilities, and permissible harm*. Oxford University Press

Kant, I., 1785. *Kant: groundwork of the metaphysics of morals*, 2nd ed, Korsgaard, C. (Ed.), 2012. Translated by M. Gregor and J. Timmermann. Cambridge University Press

Kaufman, A.S., 1966. Practical decision. *Mind*, 75(297), pp. 25-44.

Kearns, M. and Roth, A. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press

Khoury, A.C., 2012. Responsibility, tracing, and consequences. *Canadian Journal of Philosophy*, *42*(3-4), pp. 187-207

Khoury, A.C. 2013. Synchronic and diachronic responsibility. *Philosophical Studies*, 165(3), pp. 735-752

King, M., 2009. The problem with negligence. *Social Theory and Practice*, 35(4), pp.577-595

Kroes, P., 2016. Experiments on socio-technical systems: the problem of control. *Science and Engineering Ethic*s, 22(3), pp. 633-645

Larson, J., Mattu, S., Kirchner, L. and Angwin, J., 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica*, 23 May 2016. Available at https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [Accessed 29 June 2021]

Latour, B., 1992. Where are the missing masses? the sociology of a few mundane artefacts. In Bijker W. & Law J., 1994, *Shaping technology / building society: studies in sociotechnical change*. MIT Press

Latour, B., 1994. On technical mediation: philosophy, sociology, genealogy. *Common Knowledge*, 94(4), pp. 29–64

Latour, B. 2005. *Reassembling the social: An introduction to actor-network-theory*. Oxford University press.

Law Commission, 2018. Automated vehicles: a joint preliminary consultation paper. *Law Commission Consultation Paper no. 240*. Available at *https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2018/11/6.5066_LC_AV-Consultation-Paper-5-November_061118_WEB-1.pdf* [Accessed 25 June 2021]

Law Commission, 2019. Automated vehicles: consultation paper 2 on passenger services and public transport. *Law Commission Consultation Paper no. 245* Available at https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2019/10/Automated-Vehicles-Consultation-Paper-final.pdf [Accessed 25 June 2021]

Law Commission, 2020. Automated vehicles: consultation paper 3 – a regulatory framework for automated vehicles. *Law Commission Consultation Paper no. 252*. Available at https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2021/01/AV-CP3.pdf [Accessed 25 June 2021]

LeCun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. *Nature*, 521(7553), pp. 436-444

Levy, N., 2005. The good, the bad and the blameworthy. *Journal of Ethics & Social Philosophy*, 1(2), pp. 1-15

Lin, P., Abney, A. & Bekey, G.A. (Eds), 2012. *Robot ethics: the ethical and social implications of robotics*. MIT Press

Lin, P., 2014. Here's a terrible idea. In *Wired,* Aug 2014. Available at https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/ [Accessed 29 June 2021]

Lin, P., Jenkins R. & Abney, K. (Eds), 2017. *Robot ethics 2.0: from autonomous cars to artificial intelligence*. Oxford University Press

List, C. & Pettit, P., 2011. *Group agency: the possibility, design, and status of corporate agents*. Oxford University Press

Locke, J., 1689. *An essay concerning human understanding*, Peter H. Nidditch (Ed.), 1975. Oxford Clarendon Press

Loh, W. & Loh, J., 2017. Autonomy & responsibility in hybrid systems. In In Lin, P., Jenkins R. & Abney, K. (Eds), *Robot ethics 2.0: from autonomous cars to artificial intelligence*, pp. 35-50. Oxford University Press

Lucas, J.R., 1995. *Responsibility*. Clarendon Press.

Ludwig, K., 2014. Proxy agency in collective action. *Noûs*, 48(1), pp. 75-105

Ludwig, K., 2017. *From plural to institutional agency: collective action II*. Oxford University Press

McCall, S., 1987. Decision. *Canadian Journal of Philosophy*, 17(2), pp. 261-287

McCulloch, W.S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5, pp. 115-133

McDermid, J., Jia., Y., Porter Z. & Habli, I., 2021 (forthcoming). AI explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*

McKenna, M., 2012. *Conversation and responsibility*. Oxford University Press

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A. & Etemadi, M., 2020. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), pp. 89-94

McLaren, B., 2003. Extensionally defining principles and cases in ethics: an AI model. *Artificial Intelligence*, 150, pp. 145-181

McLaren, B., 2005. Lessons in machine ethics from the perspective of two computational models of ethical reasoning. *2005 AAAI Fall Symposium on Machine Ethics*, pp. 1-8

McMahan, J., 2010. Responsibility, permissibility, and vicarious agency. *Philosophy and Phenomenological Research*, 80(3), pp. 673-680

Matthias, A., 2004. The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3), pp. 175-183

May, L., 1987. *The morality of groups: collective responsibility, group-based harm, and corporate rights*. University of Notre Dame Press

Mele, A.R., 2017. Direct control. *Philosophical Studies* 174, pp. 275–290

Mellor, R., 2021. Faces of vicarious responsibility. *The Monist*, *104*(2), pp. 238-250

Menary, R. (Ed), 2010. *The extended mind*. MIT Press

Merat, N., Seppelt, B., Louw, T., Engström, J., Lee, J.D., Johansson, E., Green, C.A., Katazaki, S., Monk, C., Itoh, M. & McGehee, D., 2019. The "out-of-the-loop" concept in automated driving: proposed definition, measures and implications. *Cognition, Technology & Work*, 21(1), pp. 87-98

Miikkulainen, R., 2021. Creative AI through evolutionary computation: principles and examples. *SN Computer Science*. 2, 163

Millar, J., 2015. Technology as moral proxy: autonomy and paternalism by design. *IEEE Technology and Society Magazine*, 34 (2), pp. 47-55

Millar, J. & Kerr, I., 2016. Delegation, relinquishment, and responsibility: The prospect of expert robots. In Calo, R., Froomkin, A.M., & Kerr, I. (Eds), *Robot law*. Edward Elgar Publishing

Millar, J., 2017. Ethics settings for autonomous vehicles. In Lin, P., Jenkins R. & Abney, K. (Eds), *Robot ethics 2.0: from autonomous cars to artificial intelligence*, pp. 20-34. Oxford University Press

Millican, P. & Clark, A. (Eds), 1996. *Machines and thought: the legacy of Alan Turing, volume 1*. Oxford University Press

Möller, N., Hansson, S.O. & Peterson, M., 2006. Safety is more than the antonym of risk. *Journal of Applied Philosophy*, 23(4), pp. 419-432

Moore, G., 1965. Cramming more components on to integrated circuits. *Electronics magazine*, 38 (8), 19th Apr 1965

Moya, C.J., 1990. *The philosophy of action: an introduction*. Polity Press

Müller, V.C., 2020. Ethics of artificial intelligence and robotics. *Stanford Encyclopaedia of Philosophy (Summer 2020 edition)*. Available at https://plato.stanford.edu/archives/sum2020/entries/ethics-ai/ [Accessed 25 June 2021]

Muntean, I. & Howard, D.A., 2014. Artificial moral agents: creative, autonomous, social. An approach based on evolutionary computation. In Seibt J., Hakli R. & Nørskov N. (Eds.), *Sociable Robots and the Future of Social Relations, Proceedings of Robo-Philosophy Dec 2014*, pp. 217-230

Nagel, T., 1979. *Mortal questions*. Cambridge university press.

Nelkin, D.K., 2011. *Making sense of freedom and responsibility*. Oxford University Press

Nelkin, D.K. & Rickless, S.C., 2017. Moral responsibility for unwitting omissions: a new tracing view. *The Ethics and Law of Omissions*, pp. 106-129. Oxford University Press

Ney, A., 2014. *Metaphysics: an introduction*. Routledge

Nissenbaum, H., 1996. Accountability in a computerized society. *Science and Engineering Ethics*, 2(1), pp. 25-42

National Transportation Safety Board (NTSB), 2019. *Public meeting of November 19, 2019 - collision between vehicle controlled by developmental automated driving system and pedestrian in Tempe, Arizona*. Available at https://www.ntsb.gov/news/events/Documents/2019-HWY18MH010-BMG-abstract.pdf [Accessed 28 June 2021]

Nyholm, S. & Smids, J., 2016. The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical theory and moral practice* 19, pp. 1275-1289

Nyholm, S., 2018. Attributing agency to automated systems: reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, 24(4), pp. 1201-1219

OECD, 2019. *OECD principles on AI*. Available at https://www.oecd.org/going-digital/ai/principles/ [Accessed 28 Jun 2021]

Oshana, M., 2004. Moral accountability. *Philosophical Topics*, 32(1&2), pp. 255-274

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E. & Guston, D., 2013. A framework for responsible innovation. In Owen, R., Bessant, J. & Heitz, M., *Responsible innovation: managing the responsible emergence of science and innovation in society*, pp.27-50, Wiley

Responsible innovation: managing the responsible emergence of science and innovation in society. John Wiley & Sons

Pasquale, F., 2020. *New laws of robotics: defending human expertise in the age of AI*. Harvard University Press

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann

Pereboom, D., 2001. *Living without free will*. Cambridge University Press

Pettit, P., 2007. Responsibility incorporated. Ethics, 117(2), pp.171-201

Pitkin, H.F., 1967. *The concept of representation*. University of California Press

Porter, Z., Habli, I., Monkhouse, H. & Bragg, J., 2018. The moral responsibility gap and the increasing autonomy of systems. *International Conference on Computer Safety, Reliability, and Security*, pp. 487-493

Powers, T.M., 2006. *Prospects for a Kantian machine. IEEE Intelligent Systems*, 21(4), pp. 46-51

Powers, T.M. 2013. On the moral agency of computers. *Topoi*, 32(2), pp. 227-236

Purves, D., Jenkins, R. & Strawser, B. J., 2015. Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice*, 18(4), pp. 851-872

Raz, J. 2010. Responsibility and the negligence standard. *Oxford Journal of Legal Studies*, 30(1), pp. 1-18

Robichaud, P. & Wieland, J.W. (Eds.), 2017. *Responsibility: the epistemic condition*. Oxford University Press

Robinson, W. 2014. Philosophical challenges. In Frankish, K. & Ramsey, W.M. (Eds.), *The Cambridge handbook of artificial intelligence*, pp. 64-85. Cambridge University Press

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), pp.386 - 408

Roff, H.M., 2013. Responsibility, liability, and lethal autonomous robots. In Allhoff, F., Evans, N.G. and Henschke, A., *Routledge handbook of ethics and war: just war theory in the 21st century,* pp. 352-364. Routledge

Royal Academy of Engineering, 2015. *Innovation in autonomous systems*. Available at https://www.raeng.org.uk/publications/reports/innovation-in-autonomous-systems [Accessed 25 June 2021]

Royal Academy of Engineering, 2020a. *The journey to an autonomous transport system: identifying challenges across multiple modes*. Available at https://www.raeng.org.uk/publications/reports/the-journey-to-an-autonomous-transport-system [Accessed 25 June 2021]

Royal Academy of Engineering, 2020b. *Safety and ethics of autonomous systems*. Available at https://www.raeng.org.uk/publications/reports/safety-and-ethics-of-autonomous-systems [Accessed 25 June 2021]

Royal Society, 2017. *Machine learning: the power and promise of computers that learn by example*. Available at https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf [Accessed 25 June 2021]

Rubel, A., Castro, C. & Pham, A., 2019. Agency laundering and information technologies. *Ethical Theory and Moral Practice*, 22(4), pp. 1017-1041

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp. 206-215

Rumelhart, D.E., Hinton, G.E., & Williams, R.J., 1986. Learning representations by back-propagating errors. *Nature*, 323(6088), pp. 533-536

Russell, S. & Norvig, P., 2003. *Artificial intelligence: a modern approach (2nd edition).* Prentice Hall

Russell, S., 2019. *Human compatible*. Viking

SAE International, 2021. *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. Available at https://www.sae.org/standards/content/j3016_202104/ [Accessed 28 June 2021]

Santoni de Sio, F. & Van den Hoven, J., 2018. Meaningful human control over autonomous systems: a philosophical account. *Frontiers in Robotics and AI* 5, p. 1-17

Sapontzis, S.F., 1987. Everyday morality and animal rights. *Between the Species*, 3(3), p. 107-118

Scanlon, T.M., 1998. *What we owe to each other*. Harvard University Press

Scanlon, T.M., 2013. Giving desert its due. *Philosophical Explorations*, *16*(2), pp. 101-116.

Scharff, R.C. & Dusek, V., 2014. *Philosophy of technology, second edition.* Wiley

Schlosser, M., 2019. Agency. *The Stanford encyclopaedia of philosophy (winter 2019 edition)*. Available at https://plato.stanford.edu/archives/win2019/entries/agency/ [Accessed 25 June 2021]

Schulzke, M., 2013. Rethinking military gaming: America's army and its critics. *Games and Culture*, 8(2), pp. 59-76

Searle, J.R., 1980. Minds, brains, and programs. *Brain and Behavioural Sciences*, 3(3), pp. 417-424

Searle, J,R., 1990. Collective intentions and actions. In Cohen, P., Morgan, J. & Pollack, M. (Eds.), *Intentions in communication*, pp. 401–415. MIT Press

Searle, J.R., 1995. *The construction of social reality*. Simon and Schuster

Searle, J.R., 1997. *The mystery of consciousness*. New York Review of Books

Sepinwall, A.J., 2016. Corporate moral responsibility. *Philosophy Compass*, 11(1), pp.3-13

Shabo, S., 2015. More trouble with tracing. *Erkenntis*, 80, pp. 987-1011

Shapiro, P., 2006. Moral agency in other animals. *Theoretical medicine and bioethics*, 27(4), pp.357-373

Sharkey, A., 2017. Can robots be responsible moral agents? and why should we care*? Connection Science*, 29(3), pp. 210-216

Sharkey, N., 2008. Grounds for discrimination: autonomous robot weapons. *RUSI Defence Systems*, 11(2), pp. 86-89

Shoemaker, D., 2009. Responsibility and disability. *Metaphilosophy*, 40(3-4), pp. 438-461

Shoemaker, D., 2013. On criminal and moral responsibility. *Oxford Studies in Normative Ethics:* Volume 3, pp. 154-178

Shoemaker, D., 2015. *Responsibility from the margins*. Oxford University Press

Simpson, T., & Müller, V., 2016. Just war and robots' killings. *The Philosophical Quarterly* 66(263), pp. 302-322

Singer, P., 1979. *Practical ethics*. Cambridge University Press

Singer, P., 2011. *The expanding circle: ethics, evolution, and moral progress*. Princeton University Press

Smart, J.J., 1961. Free-will, praise and blame. *Mind*, 70(279), pp. 291-306

Smiley, M., 1992. *Moral responsibility and the boundaries of community*. University of Chicago Press

Smith, A.M., 2012. Attributability, answerability, and accountability: in defense of a unified account. *Ethics*, 122(3), pp. 575-589

Smith, H., 1983. Culpable ignorance. *The Philosophical Review*, 92(4), pp. 543-571

Sparrow, R. 2007. Killer robots. *Journal of Applied Philosophy*, 24(1), pp. 62-77

Stahl, B.C., 2006. Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology*, 8(4), pp. 205-213.

Stoutland, F., 2008. The ontology of social agency. *Analyse & Kritik*, 30(2), pp. 533-551

Strawson, P.F., 2003. Freedom and resentment. In Watson G. (Ed.), *Free will*, pp. 72-93. Oxford University Press

Street, S., 2006. A Darwinian dilemma for realist theories of value. *Philosophical Studies* 127(1), pp. 109-166

Sullins, J., 2011. When is a robot a moral agent? In Anderson, M. & Anderson, S.L. (Eds.), *Machine ethics*, pp. 151-161. Cambridge University Press

Sullivan, E., 2021 (forthcoming). Understanding from machine learning models. Available at https://philpapers.org/archive/SULUFM.pdf [Accessed 25 June 2021]

Sutton, R.S. & Barto, A.G., 2018. *Reinforcement learning: An introduction*. MIT Press.

Talbert, M., 2013. Unwitting wrongdoers and the role of moral disagreement in blame. *Oxford Studies in Agency and Responsibility* 1, pp. 225–45

Talbert, M., 2016. *Moral responsibility*. Polity Press

Tasioulas, J., 2019. First steps towards an ethics of robots and artificial intelligence. *Journal of Practical Ethics*, 7(1), pp 49-83

Thoma, J., 2021 (forthcoming). Risk imposition by artificial agents: the moral proxy problem. In Vöneky, S., Kellmeyer, P.,Müller, O. & Burgard W. (Eds.), *The Cambridge handbook of responsible artificial intelligence: interdisciplinary perspectives*. Cambridge University Press. Available at https://johannathoma.files.wordpress.com/2021/02/moral-proxy-problem-feb-2021.pdf [Accessed 25 June 2021]

Thompson, D.F., 1980. Moral responsibility of public officials: the problem of many hands. *The American Political Science Review*, pp. 905-916

Tiles, M. & Oberdiek, H., 1995. *Living in a technological culture*. Routledge

Todd, P., 2016. Strawson, moral responsibility, and the "order of explanation": an intervention. *Ethics*, 127(1), pp. 208-240

Torrance, S., 2011. Machine ethics and the idea of a more-than-human moral world. In Anderson, M. & Anderson, S.L. (Eds.), *Machine ethics*, pp. 115-137. Cambridge University Press

Tuomela, R., & Miller, K., 1988. We-intentions. *Philosophical Studies*, 53(3), pp. 367-389.

Turing, A.M., 1950. Computing machinery and intelligence. *Mind*, 59(236), pp. 433-460

Turner, J., 2019. *Robot rules: regulating artificial intelligence*. Springer

United Nations (UN), 2014. *Convention on certain conventional weapons*. Available at https://www.un.org/disarmament/publications/more/ccw/ [Accessed 25 June 2021]

University of Montreal, 2018. *Montreal declaration for a responsible development of artificial intelligence*. Available at https://5dcfa4bd-f73a-4de5-94d8-c010ee777609.filesusr.com/ugd/ebc3a3_506ea08298cd4f8196635545a16b071d.pdf [Accessed 25 June 2021]

Van de Poel, I., Royakkers, L. & Zwart, S.D., 2015. *Moral responsibility and the problem of many hands*. Routledge

Van Inwagen, P., 1986. *An essay on free will*. Oxford University Press

Van Wynsberghe, A. & Robbins, S., 2019. Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, 25(3), pp. 719-735

Vavova, K., 2015. Evolutionary debunking of moral realism. *Philosophy Compass*, 10(2), pp. 104-116

Verbeek, P. P. 2005. *What things do: Philosophical reflections on technology, agency, and design.*

Vargas, M., 2005. The trouble with tracing. *Midwest Studies in Philosophy*, 29, pp. 269-291

Verbeek, P.P., 2005. *What things do - philosophical reflections on technology, agency and design.* Penn State University Press

Verbeek, P.P., 2011. *Moralizing technology: understanding and designing the morality of things.* University of Chicago Press

Versenyi, L., 1974. Can robots be moral? *Ethics*, 84(3), pp. 248–259

Von Wright, G.H., 1963. *The varieties of goodness*. Routledge & Kegan Paul

Wachter-Boettcher, S., 2017. *Technically wrong: sexist apps, biased algorithms, and other threats of toxic tech.* W.W. Norton & Company

Waelbers, K., 2009. Technological delegation: responsibility for the unintended. *Science and Engineering Ethics*, 15(1), pp. 51-68

Waldrop, M.M., 1987. A question of responsibility. *AI Magazine*, 8(1), pp. 28-39

Wallach, W. & Allen, C. 2009. *Moral machines: teaching robots right from wrong.* Oxford University

Wallach, W. & Asaro, P. (Eds), 2017. *Machine ethics and robot ethics*. Routledge

Wallach, W., Franklin, S. & Allen, C., 2010. A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, 2(3), pp. 454-485

Watson, G., 1996. Two faces of responsibility. *Philosophical Topics*, 24(2), pp. 227-248

Watson, G., 2004. *Agency and answerability*. Oxford University Press

White, A.R., 1968. *The philosophy of action*. Oxford University Press

Wiener, N., 1948. *Cybernetics: or control and communication in the animal and the machine.* Martino Publishing (2013 reprint)

Williams, B., 1973. A critique of utilitarianism. In Smart, J.J.C. & Williams, B., *Utilitarianism: for & against*, pp. 77-150. Cambridge University Press

Williams, B., 1993. *Shame and necessity*. University of California Press

Winfield, A.F., Blum, C. & Liu, W., 2014. Towards an ethical robot: internal models, consequences and ethical action selection. In *Conference Towards Autonomous Robotic Systems (TAROS 2014),* pp. 85-96. Springer

Wolgast, E.H., 1992. *Ethics of an artificial person: lost responsibility in professions and organizations*. Stanford University Press

Wong, D.B., 1995. Pluralistic relativism. *Midwest Studies in Philosophy*, 20, pp. 378-399

Wong, D.B., 2009. *Natural moralities: a defense of pluralistic relativism*. Oxford University Press

Wooldridge, M.J. & Jennings, N.R., 1995. Intelligent agents: theory and practice. *The Knowledge Engineering Review*, 10(2), pp. 115-152

Wooldridge, M., 2020. *The road to conscious machines: the story of AI*. Penguin UK

Yazdanpanah, V., Gerding, E., Stein, S., Dastani, M., Jonker, C.M. & Norman, T., 2021. Responsibility research for trustworthy autonomous systems. Available at https://eprints.soton.ac.uk/447511/1/Responsibility_Research_for_Trustworthy_Autonomous_Systems.pdfn [Accessed 25 June 2021]

Zimmerman, M.J., 1986. Negligence and moral responsibility. *Noûs,* 20(2), pp.199-218

Zimmerman, M.J., 1988. *An essay on moral responsibility*. Rowman & Littlefield

Zimmerman, M.J., 2006. Moral luck: a partial map. *Canadian Journal of Philosophy*, 36(4), pp. 585-608.

Zimmermann, A., Di Rosa, E. & Kim, H., 2020. Technology can't fix algorithmic injustice. *Boston Review: A Political and Literary Forum*, 9th Jan 2020

Zimmermann, A. & Lee-Stronach, C., 2021. Proceed with caution. *Canadian Journal of Philosophy*, pp.1-20. doi:10.1017/can.2021.17