Reasons-Responsive Machine Compatibilism:

A New Pathway for Analysis of Autonomous Systems and Moral Responsibility Gaps

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Humanities

2022

Sarah Moth-Lund Christensen

School of Social Sciences

Content list

Word Count: 78122

Abstract

Matthias (2004) argues that the use of autonomous systems leads to 'moral responsibility gaps': cases involving the outputs of learning autonomous systems where seemingly no human agent is responsible. In this thesis, I investigate the notion of morally responsible autonomous system in order to propose a solution to moral responsibility gap problems. I argue that contemporary and near-future autonomous systems can be considered manipulated reasons-responsive entities, and further I conclude that one can trace the moral responsibility for autonomous systems' outputs and their immediate consequences back to their users.

The thesis centres on three pivotal ideas. The first is to treat non-malfunctioning autonomous systems as potential moral agents, as opposed to mere tools. This change in perspective allows one to raise questions about the potential status of autonomous systems, both current and future, as morally responsible entities. By adopting this perspective, one can investigate what conditions for moral responsibility autonomous systems might be able to fulfil – and which conditions they do not. The second is the concept of Machine Compatibilism. This thesis shows that current literature rejects the very idea of morally responsible autonomous systems by assuming that moral responsibility is incompatible with the systems' determined nature. Machine Compatibilism is introduced in response to this. The thesis further develops a machine-focused version of Fischer and Ravizza's (1998) reasons-responsive compatibilism as an example of a promising machine compatibilist account. The third is the identification of current and near-future autonomous systems as manipulated reasons-responsive entities. Together with the machine compatibilist account developed for earlier on, this provides a framework for analysing moral responsibility gap problems, and for bridging the moral responsibility gap.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree of qualification of this or any other university or institute of learning;

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=2442 0), in any relevant Thesis restriction declarations deposited in the University Library, the University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in the University's policy on Presentation of Theses.

Acknowledgements

As I have written this thesis, I have done so standing on the shoulders of all the wonderful people in my life. As such, I am grateful for this opportunity to extend a thank you to everyone, who has helped make this thesis a reality.

First, I would like to express my deepest gratitude to my supervisor and mentor, Prof. Helen Beebee. This thesis would not have been here without her invaluable insights and expert knowledge. Further, I would like to thank Prof. Beebee for her priceless support, her brilliant advice – always accompanied by generous cups of coffee – as well as her patience and great sense of humour in respect to my ever-continuing battles with the English language.

Second, a great thank you to Dr. Ann Whittle and Dr. Graham Stevens, who both have been kind enough to be part of my supervisory team, and whose comments and support helped shape this thesis.

Last, but not least, I would like to thank my partner, my friends, and my family, whose immense support and faith in me has carried me through this process. You all continue to inspire me every day.

Introduction

Machines and systems are tools and behind their outputs, use and design, humans can always be found. Yet, recent developments in the field of learning autonomous systems challenges this basic instrumental view of machines. In recent years development of autonomous systems, meaning systems capable of creating outputs without human input, has skyrocketed leaving both their creators and users seemingly without clear control of these systems' individual outputs. Matthias (2004) argues that the use of such systems leads to 'moral responsibility gaps': cases involving the outputs of learning autonomous systems, where seemingly no human agent is responsible.

In this thesis, I investigate the notion of a morally responsible autonomous system in order to propose a solution to the so-called 'moral responsibility gap' problems. I argue that contemporary and near-future autonomous systems can be considered manipulated reasons-responsive entities, and further argue that one can trace the moral responsibility for these systems' individual outputs and their immediate consequences back to their users. To do so, this thesis is divided into six chapters.

In Chapter 1, I introduce Mattias' (2004) notion of moral responsibility gaps. I argue that these present a philosophical problem: when non-malfunctioning autonomous systems provide an output that harms human beings, there is no one intuitively morally responsible for the harm in question. In this first chapter, I present examples of moral responsibility gap problems in practice and argue that current proposed solutions are severely lacking in justification – thereby leaving the philosophical problem posed by these moral responsibility gaps untreated. Two groups of human agents are usually identified as the potential morally responsible parties in scenarios featuring moral responsibility gaps: the creators and the users. I discuss current attempts at attributing the responsibility to either party and argue that current accounts are too underdeveloped to give a

satisfactory answer. In this way, the first chapter lays the foundation for the rest of the thesis, wherein I seek to answer the overarching question: who is the morally responsible party in cases featuring an alleged moral responsibility gap?

In Chapter 2, I start investigating the possibility of morally responsible autonomous systems to force a new perspective on analyses of moral responsibility gap problems. I argue that a common objection to the claim that machines can be responsible is grounded in the worry that these machines or systems do not have the freedom that is intuitively required for the attribution of moral responsibility. I present the most coherent wording of this objection as promoted by Bringsjord (2008). I then go on to argue that Bringsjord's argument is nothing more than the well-known argument for incompatibilism, the Consequence Argument, in disguise. An answer to Bringsjord's objection might therefore be found by deploying a classic compatibilist response to the Consequence Argument. I present and discuss three different compatibilist argumentative strategies for responding to the classic Consequence Argument. Due to the major influence of the Consequence Argument, any successful compatibilist account of moral responsibility would do well to utilise one of the discussed strategies – or similar – to respond to the argument. I, therefore, argue that in response to Bringsjord and similar 'machine incompatibilists', one might be able to extend or further develop an existing compatibilist account to cover autonomous systems and thereby give a positive story of how autonomous systems can fulfil the freedom-relevant condition of moral responsibility. The overarching project of the thesis therefore reveals itself: to develop a 'machine compatibilist' account for analysing moral responsibility gap problems.

By 'machine compatibilist' account, I mean an account of moral responsibility that does not render it impossible for current or near-future machines – and in particular autonomous systems – to be bearers of moral responsibility. By developing a machine compatibilist account in

this thesis, I seek to use it as a skeleton key and put it to work in multiple ways. Traditionally, machines have been seen as mere tools, and this instrumentalist view has carried on, even in analysis of cases featuring sophisticated systems, such as contemporary and near-future autonomous systems. Indeed, it is this instrumentalist view on autonomous systems that allows moral responsibility gaps to flourish, as this view renders autonomous systems nothing more than unpredictable, unwieldable, and strange tools outside of the human control we have come to expect. As such, pursuing a machine compatibilist account is first and foremost a break with this instrumentalist reading: I treat autonomous systems not as sophisticated tools, but instead as potential bearers of moral responsibility. By doing this, it should be possible to find out *why* – if autonomous systems are in fact not morally responsible for their own outputs as the machine incompatibilists claims – they are failing to be so. Hence, the development of a machine compatibilist account is not only done in order to provide an answer to moral responsibility gap cases – though that would be a fine bonus – but fundamentally to change the very way autonomous systems are being perceived in the current literature. However, this endeavour requires some practicalities to be considered; first, what kind of account of moral responsibility might in principle apply to autonomous systems, and, second, why current and near-future autonomous systems might fail to meet the requirements of that account.

In Chapter 3, I investigate the prospects for creating a machine compatibilist account based on the compatibilist theories of moral responsibility put forward by Wolf (1987) and Strawson (1962). I will argue that neither account can intuitively be extended to cover autonomous systems as potential morally responsible agents. I will argue that Wolf's account lends itself poorly to the analysis of autonomous systems as her account requires agents to have a specific mental makeup, including a 'higher self', as well as requiring fulfilment of a sanity condition. I shall then present Strawson's (1962) influential account. I will argue that there are multiple intuitive reasons

why Strawson's account cannot be easily extended to cover autonomous systems. I will in particular investigate Strawson's use of the 'reversal thesis' and his account's use of emotions, and argue that these do not allow the possibility of considering contemporary autonomous systems to be potentially morally responsible. By the end of this chapter, I will therefore have significantly narrowed down the field of potential usable compatibilist theories for my machine compatibilist project.

In Chapter 4, I present Fischer and Ravizza's (1998) reasons-responsive compatibilist account. I intend to use Fischer and Ravizza's theory as the foundation for my own machine compatibilist account, hence I use this chapter to go through all the key concepts of the theory in detail. I will show how Fischer and Ravizza argue that the notion of guidance control is necessary and sufficient for the freedom-relevant aspect requirement on for moral responsibility. As guidance control in turn requires moderate reasons-responsiveness and mechanism ownership, I shall go over each of these. I will show that reasons-responsiveness is made up of reasons-receptivity and reasons-reactivity. I will then show how Fischer and Ravizza introduce the condition of mechanism ownership due to worries about manipulation scenarios. I argue that this condition is unsuccessful in answering manipulation cases. I therefore dismiss the condition of mechanism ownership as necessary for guidance control. Thus, a different condition that deals effectively with manipulation cases is needed.

In Chapter 5, I argue that contemporary and near-future autonomous systems can be ascribed moderate reasons-responsiveness with respect to their outputs. I will argue that through analysis of the inputs and outputs of autonomous systems alone, even the simplest system can be found to be strongly reasons-reactive and exhibit regular reasons-receptivity. I will spend the last part of the chapter discussing how 'reasons' and reasons-grounded action might be understood

when analysing autonomous systems. I will here use Dennett's (1987) 'intentional stance' to provide a framework for understanding reasons-grounded behaviour in these systems.

In Chapter 6, I argue that contemporary and near-future autonomous systems qualify as manipulated agents, and I argue that their status as such can be used to provide a solution to the moral responsibility gap. In this chapter, I focus on the manipulation worries first introduced in Chapter 4. I argue that contemporary and near-future autonomous systems, while exhibiting regular reasons-receptivity, have an incredibly limited range of reasons to which they might be receptive, in contrast to human agents. I show that this limitation of reasons-receptivity in autonomous systems matches with that of indoctrinated human agents; in other words, the way in which indoctrination manipulates human agents is via limiting the range of reasons to which they are receptive. I argue that autonomous systems should therefore be seen as manipulated reasons-responsive entities. Traditionally, manipulated agents are not considered morally responsible for their actions; and indeed, as discussed in Chapter 4, it is for this reason that Fischer and Ravizza introduce mechanism ownership as a condition on guidance control, distinct from moderate reasons-responsiveness. My claim is that, while current autonomous systems satisfy the moderate reasons-responsiveness requirement, guidance control requires the agent to not be manipulated – and autonomous systems, both current and near-future, fails this requirement due to their lack of receptivity to a sufficiently wide range of reasons.

I go on to consider the question of who *is* responsible for a manipulated agent's actions. I argue that in cases featuring human agents, the manipulator can be found to be the morally responsible party, even if they do not have full foresight of what might happen as a consequence of their manipulation. I will argue that in cases featuring autonomous systems, the manipulator role is co-played by the creators as well as the users of the system. However, I will

argue that the user can be identified as the one morally responsible for the specific outputs of their given autonomous system.

Having established that autonomous systems are manipulated reasons-responsive entities, I analyse moral responsibility gap cases and identify the user as the morally responsible party for the outputs in question. Hence, I use the machine compatibilist account created in this thesis to identify autonomous systems as reasons-responsive albeit manipulated entities, thereby successfully completely diverting from the usual reading of these systems as mere tools. Having cast a new light on the nature of these systems and their outputs, the use of my machine compatibilist account allows not just for a new perspective when analysing of moral responsibility gap cases, but indeed furthermore – by identifying a morally culpable agent in these cases – it provides a means to close them for good.

Chapter 1: Mind the Gap

In this chapter I will argue that well-functioning autonomous artificial systems that work as intended pose a philosophical problem featuring so-called 'moral responsibility gaps'. I will argue that there is a current gap in the literature as well, since no satisfactory solution to the moral responsibility gap problem has been identified. To do so, this chapter is divided into three sections.

In §I, I will introduce the moral responsibility gap as formulated by Matthias (2004). I will present the original moral responsibility gap as a puzzle used to describe cases where learning autonomous systems harm one or more people due to an unforeseen interaction with its environment, yet it seems that no human agent is responsible. Since it is usually taken for granted that the learning autonomous system itself cannot be morally responsible for its outputs, this opens up a 'gap' in the attribution of moral responsibility: a poor outcome for which nobody is morally accountable. I will then argue that one can distinguish between three different types of cases in which autonomous systems harm humans. I will show that it is one of these types in particular, named 'decision-making cases', that is of philosophical interest.

In §§II-III, I will examine how the current literature on autonomous systems has attempted to address the problem of moral responsibility gaps. In cases featuring an apparent moral responsibility gap, there are usually three parties involved: the creator(s), the users and the autonomous system itself. Most literature in the field has sought to assign the moral responsibility to one of the first two parties. In §II, I will discuss the possibility of creator responsibility, and argue that the current literature has been unsuccessful in solving the moral responsibility gap by appealing to creator responsibility. In §III, I will discuss and criticise current attempts to place the moral responsibility on the users of autonomous systems. I will similarly conclude that current

theories in this regard are insufficient to address the philosophical question posed by moral responsibility gaps.

By the end of this chapter, I will therefore have introduced the core puzzle of this thesis, namely the moral responsibility gap problems. I will further have shown that outright attempts at attributing moral responsibility to one of the involved parties have been either unsuccessful or just outright dismissed in current literature. As such, this chapter lays the foundation for the following chapters, wherein I shall seek to answer the overarching question: who is the morally responsible party in cases featuring an alleged moral responsibility gap?

I. The Curious Case of Moral Responsibility Gaps

In this section, I will present the original moral responsibility gap problem as formulated by Matthias (2004), which has set the tone for most of the contemporary literature surrounding autonomous systems and moral responsibility. I will clarify the notion of 'autonomous systems' and make a linguistic road map for the use of the term 'autonomous'. I will then end this section by distinguishing between three different types of cases in which autonomous systems cause harm to human agents, and I will argue that moral responsibility gaps only supposedly arise in two out of these three types of cases.

A moral responsibility gap, as originally explained by Matthias (2004), arises in cases involving the actions of learning autonomous systems, where seemingly no human agent is responsible. To illustrate, Matthias (2004) asks the reader to imagine the following scenario. Consider an autonomous system in the form of a children's toy, developed to fit in small apartments or the homes of busy people. This toy, designed to offer companionship as a robotic pet, has the

ability to adapt its behaviour to its environment and act without human intervention (Matthias 2004, 177). It just so happens that this environment features heavy carpeting, which makes walking impossible for the robotic pet.[1] So, the pet raises its speed to fulfil its task of walking across the carpet. Unfortunately, as the robotic pet has just started to run around the apartment at high speed, it collides with a child, who happens to be in its way. The child is harmed.

Now one might then ask who is morally responsible for the harm of the child. Matthias (2004) argues that the difficulty in answering the question presents itself when considering the role of the autonomous system. The behaviour of learning autonomous systems is not defined purely by the system's original programming (Matthias 2004, 182). Instead, the system can process data from its surroundings, draw new conclusions and change behaviour accordingly. This means that for a learning autonomous system in use, the programming that determines the system's action in a given situation might not have been written by the system's initial programmer. Matthias shows that for learning autonomous systems based in logic-oriented programming or neural networks, the programmer loses their traditional control over the execution flow of the system's program (2004, 181).

Matthias provides multiple examples of similar cases featuring learning autonomous systems (2004, 176-177). By the time of writing, the literature on machine ethics has already produced an array of scenarios illustrating such gaps featuring everything from robotic pets to lethal autonomous weapon systems.[2] The key for all proposed examples is the fact that autonomous systems can produce unforeseeable outputs that in theory may result in harm to human beings. Due

[1] See Hornby et al. 2005 for more information about AIBO, the real-life robotic pet and its abilities.

[2] For examples, see Matthias 2004 for scenarios featuring pet robots and systems used for medical diagnosis. See Sparrow 2007 for a debate-defining scenario featuring lethal autonomous weapon systems, Hevelke and Nida-Rümelin 2015 for cases featuring autonomous vehicles and Gunkel 2017 for examples featuring current and past technology.

to the harm's unforeseeable nature, Matthias argues that a scenario like that above leaves a moral responsibility gap. In other words, in cases such as the one above it seems that no human can rightly be blamed for the harm caused by the output of an autonomous system (Matthias 2004, 177).

Before moving on, I will briefly make a clarification on the use of the term 'autonomous' and how it is used in debates on autonomous systems. As this is a heavily ambiguous term used haphazardly in the contemporary literature on autonomous systems, a distinction is worth making. In this thesis, I will distinguish between the following three possible meanings of 'autonomous'. In chapter 5, I will expand on the topic and distinguish between different levels of autonomy when it comes to autonomous systems, however for the time being the following brief distinctions will do. First is 'autonomous' in a philosophical sense. As with every concept in philosophy, the definition of autonomy is heavily debated. For now, I shall therefore merely use a summary definition as given in the Stanford Encyclopedia of Philosophy's entry on autonomy: "Individual autonomy is an idea that is generally understood to refer to the capacity to be one's own person, to live one's life according to reasons and motives that are taken as one's own and not the product of manipulative or distorting external forces" (Christman 2018, §0).

Imagine as an example a student, who enters their university library. If they have chosen to do so because they have a motivation and desire to study, then it might be said that they autonomously decided to go to the library. By contrast, imagine that the student walked to campus solely because they had been hypnotised: that they had been steered to the library like a mere puppet, and did not act based on any of their own reasons. The student would then not be considered as having acted autonomously in a philosophical sense. This summary in its simplicity does not reflect in any way the copious amounts of philosophical work on the concept of

autonomy.[3] However, this definition will do for the purpose of differentiating it from the following two uses of the term.

Second is 'autonomous' in the sense of folk interpretation. The criteria for this sense of autonomy are neither well-defined nor strict like those found in the philosophical tradition. The idea of autonomous systems in this sense conjures up ideas of conscious, emotional Asimovian machines, as found in films like 'I, Robot' (2004).[4] Due to this interpretation's link to pop culture's portrayal of 'autonomous' machines, 'autonomous' in this sense tends to be linked with ideas of conscious super-intelligent machines. The systems that are described as autonomous in this sense are usually eerily human-like. They are also at the time of writing completely fictional.[5]

Third and last is 'autonomous' as used in computer science. To be labelled autonomous in this sense, a system merely needs to be able to act independently of human interaction. In other words, a system being autonomous roughly means that the system can create an output with no human involved in the system's decision-making process. This sense of 'autonomy' is consistent with Matthias' use of the term in descriptions of autonomous systems (2004: 176-177).[6] It must therefore be noted that the term 'autonomy' in this sense therefore applies much more liberally than in standard philosophical contexts. Consider as an example a Roomba, a robotic vacuum cleaner. A Roomba might be able to change course, when hitting a wall or a piece of furniture. An advanced Roomba might be imagined deciding on its course of hoovering based on

[3] See Frankfurt 1971 and Bratman 2007 for examples of philosophical accounts of autonomy. See Christman 2018 for an overview of how the philosophical concept of autonomy plays into contemporary moral philosophy.

[4] In popular media, such 'overly' intelligent machines are often also used as villains. See as an example Hal 9000 in 2001: A Space Odyssey (Clarke 1968) or GLaDOS in Portal (2007).

[5] See Lucas 2013 for a discussion of how these ideas of autonomous systems are clouding the current debate.

[6] See also Roff 2013 and Lucas 2013 for the use of this sense of 'autonomy' in the debate surrounding autonomous systems and moral responsibility.

the available data of the room layout. That would constitute creating an output without human interference. This advanced Roomba would therefore qualify as being autonomous in a computer science sense.

Due to the major disparities in how the different senses of 'autonomy' may be applied, it is possible for certain entities to be considered autonomous in one sense of the word, but not in another. As an example, the abovementioned Roomba cannot be understood to live its own life following reasons it has taken as its own. It vacuums because it is fundamentally programmed to do so. Hence, a Roomba is not 'autonomous' in the philosophical sense. Similarly, the Roomba fails to live up to the folk interpretation of 'autonomous'. A Roomba, even an advanced one, bears no trace of having any of the idealised traits of futuristic machines, such as consciousness. A system labelled 'autonomous' in one of the less demanding senses should therefore not be assumed to be 'autonomous' in a more demanding sense.

As will be further shown and discussed in chapters 5 and 6 in this thesis, all three senses of 'autonomous' crop up in the existing literature on autonomous systems. To avoid confusion, it should be noted that the term 'autonomous' in 'autonomous system' will in this thesis exclusively refer to autonomy in the third sense unless clearly stated otherwise. The technology needed for autonomous machines in a folk sense is still far from existing, and where the current technology stands in terms of philosophical autonomy is a question, I will leave for chapter 5.

To end this section, I will in argue that Matthias' description of cases featuring moral responsibility gaps may cover three different types of cases, where the output of autonomous case may result in harm to human beings. I will make the distinction between these three in order to clarify the scope of this thesis.

I will here use the example of self-driving cars to illustrate these types of cases. Google, Tesla and Ford are just a few of the companies that are currently selling or testing self-driving cars, while racing to make the technology increasingly independent of humans. Yet cars crash. In Arizona, on the 18th of March 2018, a pedestrian died following a collision with one of Uber's self-driving cars (Levin et al. 2018). While it has been argued that self-driving cars will on average make the streets safer, it must still be assumed inevitable that car crashes with self-driving cars will happen.[7]

An autonomous system in the form of a self-driving car is not created by a single person, but most likely by a large research team each doing their part. For example, a group of material scientists might be developing the best kind of material the car should be made of, a team of geographers are working on the perfecting the car's GPS (global positioning system), and so on. In a similar manner, one can assume that a team of software programmers have worked on the car's decision-making abilities. As a more specific example, consider one of Google's self-driving cars. The 'self-driving' aspect of the Google car is a mix of multiple engineering efforts, including the navigating autonomous system, radar sensors, GPS, laser range finders and more. In other words, a self-driving car might be viewed as an autonomous decision-making system communicating with a whole lot of 'gadgets', which constantly feed the system with information about its current environment.[8]

Imagine that a human agent is using their autonomous car to get from point A to B. At the time of writing, passengers in self-driving cars are still advised to sit behind the wheel so that

[7] See Marchant and Lindor 2012 for a legal discussion leading to the assumption of this inevitability. See also Schoettle and Sivak 2015, van Loon and Martens 2015, and Yang et al. 2016 for analyses and discussion of autonomous vehicle crashes in traffic featuring both autonomous as well as manually driven vehicles.

[8] See Gibbs 2014 and Birdsall 2014 for light introductions to how Google's self-driving cars work.

they may intervene if they find it necessary. Despite this, it has been shown that the experience of being in a self-driving car is phenomenologically like that of being an ordinary passenger in a car, rather than being the driver (Coeckelburgh 2016, 754). As such, for the sake of argument, let us imagine this particular car of near-future origin, where the user of the car merely enters the car and inputs the destination point. In this way, the human agent may be imagined to be purely a passenger.

Suppose that the car turns a corner and detects a child in the middle of the road. Imagine that the car's sensors fail to register the child's presence on the road. The car therefore does not alter its course and the child is run over. The harm is here caused by a malfunction of the system's sensors. For the purposes of this thesis, I will refer to this type of case as a 'malfunction case'.

Contrast the malfunction case with the following imagined scenario featuring another collision. Suppose that the vehicle's autonomous decision-making system is built using deep learning. In colloquial terms, one might imagine the system as having been initially programmed with basic abilities like the ability to drive forward, turn to either side and brake. The system has then been trained in a set of computer simulations with varying set goals, followed by an evaluation of its ability to solve the tasks given to it. Through the training, the system might be imagined learning not to hit the brakes hard when the road is wet. It might also have learned crash optimisation. Crash optimisation here refers to the ability to choose what to collide with when a collision is unavoidable.[9] This could, for example, be choosing to collide with a truck instead of a motorcycle, as the truck's passengers are less likely to be harmed grievously in a car collision than a

[9] For more discussion of the development and use of crash optimisation, see Gurney 2016.

motorcyclist.[10] Note that after the initial programming, the system is a black box to the programmers.[11]

The black box aspect deserves further clarification. A learning autonomous system is commonly equipped with some initial programming, allowing it to learn from its surroundings. This means that it will create new outputs based on its learned data. The longer such a system is in use, the more the system will create outputs that are completely based on its own learnings, thereby seceding itself from its creators. As such, these systems are more than capable of producing outputs that are completely unforeseen or even unwanted by its creators. When in use, the creators have no control over the system's individual outputs, nor insight in why these come about – hence the system becomes a black box to its creators.

Let us return to the self-driving car discussion. Suppose further that the programmers have tested the system in a variety of training simulations. The autonomous system continues to learn and continuously optimises its algorithm, while being tested in the real world as well. Imagine now that one day the self-driving car is out in heavy rain, so the roads are wet and slippery. Suppose that the car has learned that in such conditions swerving to avoid collisions is more effective than attempting to brake. To make matters worse, suppose that the car is trapped in a situation where it must choose between colliding with a big truck or swerving into a pedestrian bystander. The autonomous system acts based on the data available to the system at the time, including both information about its environment and its learned information from previous outings. It swerves into the pedestrian, killing them in an instant. For the purpose of this chapter, this type of case will be referred to as a 'decision-making' case.

[10] See Lin 2016 for an ethical analysis of crash optimisation in self-driving cars.
[11] See Das 2008 for an introduction to programming of decision-making agents.

By this point, three different types of cases featuring autonomous systems have been presented. The first - the pet toy case - was the 'accidental case', which Matthias (2004) uses to argue for a moral responsibility gap. The second is the 'malfunctioning' case and third is the 'decision-making' case. One should be careful not to mistake the 'accidental' case and the 'decision-making case' for equals. In the accidental case, the output of the autonomous system in question consisted of the robotic dog running. The harm of the child was a mere unintended by-product of this output - a case of bad luck.

The decision-making case featuring the autonomous car was unlucky in a sense as well. However, the system's output was based on crash optimisation. The system produced an output that did not just consist of swerving, but specifically swerving into the pedestrian. As such the output itself constituted the harm.

The difference might be better seen in analogous cases featuring human agents. An accident case might be likened to an agent, who decides to run around a room, but in their path accidentally knocks over a vase. In contrast, consider a human agent, who trips and can only decide which way to fall. Either they will fall into a vase, or into an irreplaceable painting. The agent chooses to fall towards the vase, which upon contact immediately shatters. The latter example is analogous to decision-making cases. The difference is simply that in decision-making cases the harm is not a by-product of poor luck, but instead the harmful action is specifically chosen. I will in this thesis primarily focus on moral responsibility gaps found in decision-making cases and occasionally accidental type cases. Therefore, the question of who is morally responsible for malfunctioning autonomous systems is left for another writer to take up.

In this section, I have introduced the moral responsibility gap problems posed by the development of autonomous systems, and clarified the notion of autonomy in relation to these. I

have further shown that one may distinguish between three types of cases in which autonomous systems cause harm to human agents. In the following sections, I will use this distinction to discuss current attempts at placing moral responsibility in the accidental and decision-making type cases.

II. Creator Responsibility and Autonomous Systems

In this section, I will present current discussions concerning the possibility that the creators of autonomous systems are morally responsible for the outputs of the systems. At the time of writing, there are, to the best of my knowledge, no authors who have argued in favour of creators being morally responsible for the outputs of their creations, so I will here present and discuss two arguments against creator responsibility. As will be shown in this section, this area of literature is severely lacking in material and the arguments that will be presented here reflect that. I will conclude that more research is needed before it is possible to make a strong argument about the moral responsibility of creators.

Before starting, it should be said that outside of Matthias' lone paper, there is little written on the topic of general learning autonomous systems and moral responsibility. Instead, papers tend to focus on worries surrounding specific autonomous systems. I will draw on some of these debates here, specifically papers from discussions of autonomous systems used as part of self-driving vehicles and lethal autonomous weapon systems. It should be noted that my distinction between the different senses of 'autonomy' is not a commonly used; the sense of 'autonomous' changes from paper to paper. This lack of distinction also makes it unclear whether the autonomous systems being discussed are reflective of current technology or mere science fiction speculation. To avoid any such ambiguity here, I will here state that this thesis will be discussing solely learning autonomous systems in a computer-science sense of the word.

In this thesis, most of my examples will feature autonomous systems, which are at a very high level of autonomy. I will say more about the different levels of autonomy in chapter 5. However, in practice this means that many of the examples will feature autonomous systems that are more sophisticated than the ones currently commercially available, such as the ones found in the current self-driving car industry. This thesis seeks to address moral responsibility gap problems, which may be posed by not only contemporary autonomous, but also near-future systems currently in development. An example of such systems may be found in this description by the United States Department of Defence (2011) in their 'Unmanned Systems Integrated Roadmap FY 2011-2036': "To operate in complex and uncertain environments, the autonomous system must be able to sense and understand the environment… The perception system must be able to perceive and infer the state of the environment from limited information… This understanding is needed to provide future autonomous systems with the flexibility and adaptability for planning and executing missions in a complex, dynamic world."[12]

I will, however, follow the trend in the current emerging literature by discussing cases featuring self-driving cars and lethal autonomous weapon systems. In this section, I will start by presenting the possibility of holding creators responsible for the actions of lethal autonomous weapon systems. The discussion of lethal autonomous weapon systems has its roots in the area of war ethics. In a highly influential paper, Sparrow (2007) warns about the ethical concerns surrounding the use of lethal autonomous weapon systems (LAWS) in both modern and future warfare. Sparrow identifies a responsibility gap in cases where LAWS commit war crimes, yet no human agent can seemingly be rightly blamed (2007, 66). His paper concludes that it would be

[12] See also Weiss 2011 for a further discussion of current developments in autonomous systems destined for use in war.

unethical to use LAWS in active warfare, as no one can justly be held directly responsible for the specific harms that LAWS might cause to other human beings.

Sparrow's descriptions of lethal autonomous weapon systems are worth briefly discussing. Consider the following passage, in which Sparrow discusses potential reasons why a lethal autonomous weapon system might commit a war crime by killing surrendering enemy soldiers: "…Perhaps it killed them because it calculated that the military costs of watching over them and keeping them prisoner were too high, perhaps to strike fear into the hearts of onlooking combatants, perhaps to test its weapon systems, or because the robot was seeking to revenge [sic] the 'deaths' of robot comrades recently destroyed in battle." (Sparrow 2007, 66). On the same page, he even refers to lethal autonomous weapon systems as 'robot warriors'. Recall from the previous section that a distinction was drawn between three uses of the term 'autonomous'. Sparrow's descriptions of the possible decision-making scenarios for LAWS suggests that he borders on using 'autonomous' in a folk sense.

Sparrow's description of lethal autonomous weapon system as autonomous in a folk sense has caused critics to dismiss his worries about responsibility and LAWS. As an example, Lucas (2013) accuses Sparrow and writers inspired by him of mistakenly assuming LAWS to be pop-cultural images of killer robots.[13] Lucas describes mockingly the current image of LAWS as follows: "R2D2 and C3PO, fully weaponised and roaming the mountains of southern Afghanistan, but unable to distinguish (without human supervision) between an enemy insurgent and a local shepherd." (2013, 8). Nevertheless, I will suggest that despite Sparrow's use of 'autonomous' in a

[13] See as an example also Arkin 2010, who envision LAWS as performing better ethically on the battlefield than human soldiers.

folk sense, his brief discussion of creators in regard to the placement of moral responsibility is still relevant when considering cases featuring autonomous systems in a computer science sense.

Sparrow argues that creators cannot be held responsible for the actions of their creations (2007, 70). Sparrow's example above of a case in which LAWS harm a human being is clearly a decision-making case. In his brief discussion, Sparrow therefore makes sure to note that the harm is not caused by negligence of the programmers (2007, 69). Instead, he notes that the more autonomous a system is, the more likely it is to make choices that could not have been foreseen by the programmers. The usual connection between the actions of a system and its creator, is therefore broken by the autonomous state of the system.

Sparrow therefore denies that creators should be held responsible, while making the following analogy: "To hold the programmers responsible for the actions of their creation, once it is autonomous, would be analogous to holding parents responsible for the actions of their children once they have left their care." (Sparrow 2007, 70). As this sentence is all Sparrow writes on the topic, the argument can be considered rather sparse. Furthermore, Sparrow's one-sentence argument does not address concepts like parental neglect, which analogously could be used as an intuitive argument in favour of holding creators responsible.

I will therefore expand on Sparrow's analogy. Consider the following case. Imagine a set of parents bringing up a child named Mark with the explicit purpose of him one day joining the army. The child through its upbringing attends a long list of extra-curricular activities in preparation for their enlisting. Eventually Mark is old enough, gets his military training, and is eventually sent to the frontier. Suppose that Mark has been in the field for a couple of years and has learned so many new things based on his experience on the frontier. One day, while on patrol, a conflict escalates and Mark makes the decision to shoot another agent, who succumbs to the gun wounds.

As Sparrow implies, Mark's parents cannot be rightly considered morally responsible for Mark's decision to shoot in that moment. Mark's parents, we may suppose, could not have foreseen that Mark would have ended up in that situation, and they never taught him how to deal with such specific circumstances, but instead he learned what to do during his time on the frontier. While Mark's parents did pave the way for Mark's military training, it was Mark's decision to shoot at that point in time, not his parents'. It would therefore seem wrong to hold Mark's parents responsible for the death of the other agent, as they were not directly connected to the decision causing the death. In other words, the parents might have paved the way and given Mark the abilities necessary for military combat, but they did not make the decision to pull the trigger in this case. Unless the parents have a story of neglect that would have caused Mark to shoot in that instance, it seems wrong to hold them responsible for the death of the other agent.

Sparrow's own argument is scant, yet the 'Mark case' shows that the argument does hold some intuitive weight in scenarios potentially analogous to cases featuring lethal autonomous weapon systems. While Sparrow himself seems to be operating with the notion of 'autonomous' systems in the folk sense, his argument, as scant as it is, does equally apply to 'autonomous' systems in a computer science sense in decision-making cases. Sparrow identifies the moral responsibility gap as being a problem solely due to the systems' decision-making abilities. This is a feature of 'autonomous' systems in a computer science sense, as well as Sparrow's imaginary 'killer robots'.

There are two different discussions in play here. The first is just whether Lucas (2013) can dismiss Sparrow's worries about lethal autonomous weapon systems based purely on the fact that Sparrow seems to understand autonomous systems to be autonomous in a folk sense. Based on what has been shown here, I argue that he cannot. Sparrow's core argument is just that creators

cannot be morally responsible for their system's actions, as the system seems to act and decide at its own behest. As autonomous systems, who are autonomous in the computer science sense, seem to do the same, Lucas' (2013) criticism falls short.

The other part of the discussion is then whether Sparrow successfully shows that creators cannot be held morally responsible for the outputs of their creation. I will just suggest here that Sparrow's argument is far too underdeveloped to answer this question. Sparrow's argument, as has been shown, is one by analogy – it assumes that the creator/LAWS relationship is analogous to a normal parent/child relationship. However, Sparrow gives us no reason to agree to such an assumption. Further, even if one assumed that such an analogy could be justifiably made, one might still use considerations of parental neglect or even manipulation to question the moral responsibility of the parents/creators. In short, the topic quickly becomes much more nuanced than Sparrow makes it out to be, and considering the fact that Sparrow dedicates less than a paragraph to this idea, calling it under-developed is more than a fair assessment.

Having presented an argument, albeit it be lacking in strength, about creator responsibility from the literature surrounding lethal autonomous weapon systems, I will now move onto an argument from the area surrounding autonomous vehicles. Early scholarship on crashes featuring self-driving cars has mostly been penned by legal scholars[14]. Focusing especially on American civil and criminal law, the early debate is therefore dominated by the discussion of legal liability in crash cases featuring autonomous vehicles. Inspired by the legal debate, Hevelke and Nida-Rümelin (2014) seek in their philosophical paper to provide an ethical analysis of the responsibility for crashes involving self-driving cars. Their analysis considers both the creators and

[14] For an introduction to the ongoing legal debate on autonomous vehicles, see Beiker 2012, Gurney 2015, Marchant and Lindor 2012.

users of autonomous vehicles as potential bearers of responsibility for car crashes with autonomous vehicles.

In the first short section of their paper, Hevelke and Nida-Rümelin investigate the possibility of holding the creators or manufacturers responsible for crashes featuring autonomous vehicles. To start, Hevelke and Nida-Rümelin note that holding creators responsible seems obvious, as the quality of an autonomous system as a finished product is dependent on the creators (Hevelke and Nida-Rümelin 2014, 620). Despite this, Hevelke and Nida-Rümelin deny that creators are responsible for crashes featuring autonomous vehicles in the first section of their paper.

Their argument is brief and pragmatic. They argue: "If the introduction of autonomous vehicles might reduce the yearly toll in death and injury exacted by road traffic even by a small degree, that would constitute a powerful moral reason in favour of promoting their development…" (Hevelke and Nida-Rümelin 2014, 623). Hevelke and Nida-Rümelin thus start their ethical analysis from a practical standpoint, by claiming that there is good moral incentive to encourage manufacturing of self-driving cars.

It is then from this point of view that they try to address the question of creator responsibility. They write: "Why should they not have to take responsibility? The clearest answer is a practical one: if in the case of crashes involving autonomous vehicles the main responsibility were to be that of the manufacturers, '*the liability burden on the manufacturer may be prohibitive of further development'* (Marchant and Lindor 2012)." (Hevelke and Nida-Rümelin 2014, 620).

Therefore, holding the creators responsible would go against the supposed strong moral reason to promote further development of autonomous vehicles. Based on this, Hevelke and Nida-Rümelin (2014) conclude that holding creators or manufacturers of autonomous vehicles morally responsible for potential crashes is not a viable nor morally right option.

This argument might strike any philosopher as curious, so a few things are worth clarifying about Hevelke and Nida-Rümelin's dismissal of creator responsibility in crash cases featuring autonomous vehicles. First is the type of responsibility discussed in their paper. Despite the paper's foundation in the legal literature, Hevelke and Nida-Rümelin are explicitly discussing responsibility from an ethical perspective and not as a question of legal liability. Hevelke and Nida-Rümelin use the concepts 'taking responsibility', 'holding responsible', 'being liable' and 'being morally responsible' interchangeably. I will in a moment show how this inconsistency is fundamentally the undoing of their argument.

Second is Hevelke and Nida-Rümelin's (2014) concept of moral responsibility and how it departs from philosophical tradition. Recall that they argue that car manufacturers cannot be morally responsible for car crashes featuring autonomous systems, as this could hurt production of autonomous cars, which has an overall positive effect on the society. To see why Hevelke and Nida-Rümelin's assumptions about attribution of moral responsibility is out of the ordinary, consider how these assumptions work in the following scenario.

Imagine that a country is ruled by a wicked king, who is set on making life as miserable as possible for all of his subjects. One day, a servant in the king's castle sees the prince poison his father's wine. The king is murdered, and the son is instated as regent. The prince is kind and brings peace and prosperity to his country.

Suppose that in this fictional country holding someone morally responsible for murder always leads to punitive measures of some kind. The servant might then have known that holding the prince responsible for his father's murder would prohibit him from being a great king. Letting the prince get away with patricide would enable him to be the new regent and end immense suffering. Based on this, the servant chooses to keep silent.

Now, the question is: is the prince morally responsible for his father's death? Using Hevelke and Nida-Rümelin's argument, the prince cannot be morally responsible for the king's death, as this would prohibit the prince from bringing great prosperity to the country.

This story shows how Hevelke and Nida-Rümelin's concept of moral responsibility diverts from standard philosophical practice. In philosophical tradition, placement of moral responsibility is dependent on factors such as an agent's worthiness of certain reactive attitudes, or whether an action can be rightly attributed to the agent in question.[15] Using such notions of moral responsibility, it is clear that the prince is morally responsible for murdering the king.

However, Hevelke and Nida-Rümelin break with such notions of moral responsibility. First, they suggest that attributing moral responsibility is equivalent to something negative like putting in place a punitive measure, which could in turn be prohibitive of some greater good, e.g. the kingdom's happiness or the lowering of car crash fatalities. Second, they assume that attribution of moral responsibility can be done selectively. In the case of the prince, there would be a great benefit or good, if no punitive measures were taken against the prince – as such, he is not morally responsible. According to Hevelke and Nida-Rümelin's account, they assume the morally right thing to do is promoting this potential good, i.e. ensuring the prince's prosperous reign. It is therefore morally right to not place any negative measurements on the prince, which is equivalent on their account to absolving the prince of moral responsibility. Hence, the prince cannot be morally responsible according to Hevelke and Nida-Rümelin's argument. It should now be clear how this account's notion of moral responsibility is unusual in comparison to standard philosophical practice.

[15] For more on reactive attitudes, see Strawson 1962, Glover 1970 and Zimmerman 2010

Hevelke and Nida-Rümelin assume that placement of moral responsibility is not dependent on appropriateness or accountability but instead dependent on whether the attribution of moral responsibility has some kind of consequential benefit. There are instrumental accounts of moral responsibility out there which Hevelke and Nida-Rümelin could use to further their argument.[16] However, this would be quite a controversial move and without at least dedicating some space to mention such a move or defend it, one is tempted believe that Hevelke and Nida-Rümelin's concept of moral responsibility as it comes across here is accidental, rather than being a calculated philosophical move.

By using the story of the prince, I have therefore now hopefully explained the assumptions that lie behind Hevelke and Nida-Rümelin's brief argument. Using these assumptions, it can explain why Hevelke and Nida-Rümelin reject the notion of car manufacturers as morally responsible. A brief summary of their argument can then be written as follows:

> P1. Production of autonomous vehicles has a great benefit to society.
>
> P2. If production of autonomous vehicles has a great benefit to society, then car manufacturers should not be held morally responsible for car crashes featuring autonomous vehicles.[17]
>
> P3. If car manufacturers should not be held morally responsible for car crashes featuring autonomous vehicles, then car manufacturers are not morally responsible for the crashes.

[16] See Vargas (forthcoming) and Jefferson 2019 for examples of instrumentalist accounts of moral responsibility.

[17] The reasoning for this premise is explained by the prince story.

C1. Car manufacturers are not morally responsible for crashes featuring autonomous vehicles.

Premise 2 was the one explained by the prince story above. I will argue that premise 2 and 3 hides some contentious claims, and that these renders Hevelke and Nida-Rümelin's argument unpersuasive. The first is a false equivocation between 'holding X morally responsible for action Y' with 'Punishing X for performing Y'. This was also highlighted by the prince story and touched upon briefly above. Hevelke and Nida-Rümelin argue that car manufacturers should be held morally responsible for their creations, as this might be prohibitive of further industry. However, traditionally when talking about moral responsibility and sanctions from one's moral community, it usually refers to disapproval or moral blame. Hence, Hevelke and Nida-Rümelin's argument is built on the claim that moral blame would stop car manufacturers from developing their technology further. While one can hope, this claim seems highly unlikely to say the least.

The second is an equivocation of the two questions: 'Are car manufacturers morally responsible for car crashes involving autonomous cars?' and 'Are car crashes featuring autonomous cars a kind of event that we, as a moral community, wish to hold someone accountable for?'. The first question is the one we are interested in, as this is the question of moral responsibility. The latter is a question of morality; its related debate could sensibly involve questions about whether the use overarching good of using autonomous vehicles means there should be no punitive measurements against the creators for crashes involving the technology.[18] The two questions can be easily seen to not be equal, nevertheless Hevelke and Nida-Rümelin seem to be more interested in the latter. This can be shown be considering what would happen on Hevelke and Nida-Rümelin's account, if the benefit of using autonomous cars did not outweigh the number of crashes they caused. In that case,

[18] The equivocation of these two questions in Hevelke and Nida-Rümeln's paper is also noted and briefly touched upon in Nyholm 2018.

according to Hevelke and Nida-Rümelin's line of thought, this would open up the possibility of holding the car manufacturers morally responsible. However, this would be an incredibly strange conclusion – because the car manufacturers and their properties, motives and control over their creations would not have changed between the scenarios. As such, if they are morally responsible in one case, they should intuitively be morally responsible in the other. Hence, Hevelke and Nida-Rümelin's argument relies on a confused equivocation of two very distinct questions.

Hevelke and Nida-Rümelin give no reasons for assuming either of these claims. I have in this section been incredibly gratuitous with my interpretation of Hevelke and Nida-Rümelin's argument, as the authors' original argument is not much more than the quotations that have been written up here as well. However, as has been shown here, even if one tries to investigate or make sense of their argument, multiple prima facie problems arise. As such, it must be fair to conclude here that Hevelke and Nida-Rümelin's argument is, as it stands, too unrefined to be found compelling.

In this section, I have presented two papers, both of which present short arguments against creator responsibility. The first presented was Sparrow's comment on creator responsibility relating to lethal autonomous weapon systems. I argued that Sparrow's intuition on creator responsibility relates to decision-making cases as well. I concluded that Sparrow's argument could be expanded to argue against creator responsibility in decision-making cases. However, as Sparrow's own argument was no more than a single line, a more nuanced account detailing the argument would be necessary before a firm rejection of creator responsibility can be made.

The second argument presented was from the literature surrounding autonomous vehicles. This paper by Hevelke and Nida-Rümelin was found to rely on a strange notion of moral responsibility, where placement of moral responsibility would be dependent on a type of utilitarian

cost-benefit analysis. I rejected Hevelke and Nida-Rümelin's brief argument based on their use of this interpretation of moral responsibility.

The two arguments here presented are the most prominent arguments in the literature currently on the topic of creator responsibility. Out of those two, Sparrow's is the strongest argument, though very underdeveloped. As such, it must be concluded that current arguments on the topic of creator responsibility are insufficient to help clear the debate surrounding moral responsibility gaps.

III. User/Implementor Responsibility and Autonomous Systems

In this section, I will present current attempts to hold implementors and users of autonomous systems responsible for the actions of the systems. I will use the term 'implementor' here interchangeably with 'user', as the appropriate term differs depending on the autonomous system in question. The difference between the two terms in the context of autonomous systems is mostly just optics. One might refer to the passenger in an autonomous car as a user of the system, where a commanding military officer might implement an autonomous system in the form of a submarine system somewhere in the field of war. For the purposes of allocation of moral responsibility, the two terms will be assumed to refer to the same group of agents.

Just as in the previous section, I will here present arguments concerning the moral responsibility gap from the literature on lethal autonomous systems and self-driving cars respectively. I will argue that there is no current satisfying answer to the question whether, in decision-making cases, users/implementors of autonomous systems might be the bearers of moral responsibility, and hence provide the solution to the moral responsibility gap problem.

To start, I will return to the debate surrounding lethal autonomous weapon systems. Champagne and Tonkens (2015) propose a solution to the moral responsibility gap in cases featuring LAWS. The proposed solution is written in response to Sparrow's aforementioned argument against the deployment of LAWS. This paper, just like Sparrow's, is therefore focused on cases where lethal autonomous weapon systems cause unforeseen harm or commit "war crimes", to use Sparrow's words (2007, 66).

Champagne and Tonkens identify and accept two core assumptions about these type of cases in Sparrow's paper (Champagne and Tonkens 2015, 134):

1. Programmers are not viable contenders for moral responsibility.
2. The autonomous system in question is not a viable contender for moral responsibility either.

Champagne and Tonkens accept these premises as true for the following main reasons. First, so they can establish that there are apparent moral responsibility gaps in these cases featuring lethal autonomous weapon systems. Second, by accepting these premises, user responsibility becomes the only available option for closing the gap. They then go on to argue that "…if a commanding officer willingly deploys autonomous robots that can act immorally and moreover publicly acknowledges that those robots cannot be held responsible, then she has thereby accepted responsibility for their actions." (Champagne and Tonkens 2015, 134).

They refer to this idea as 'blank check' responsibility and suggest it as a type of contract that can be used to avoid moral responsibility gaps in cases featuring LAWS. They write further: '…by willingly agreeing to the terms of a contract, the informed agent(s) impute(s) responsibility on herself for the actions of an autonomous machine. The missed alternative we want to highlight, then, would essentially consist in an exchange: Social prestige in the occupation of a

given office could come at the price of signing away part of one's freedoms to a contingent and unpredictable future guided by another (in this case, artificial) agency' (Champagne and Tonkens 2015, 127). The key notion behind the 'blank check' approach is then that a high-ranking commanding officer in the military could pre-emptively take on moral responsibility for potential outputs of any deployed lethal autonomous weapon systems. The officer would take on this responsibility in exchange for gaining their position and the prestige that accompanies it.

In this way, Champagne and Tonkens argue in favour of holding an implementor, which in this case refers to the commanding officer, morally responsible for the output of a given autonomous system. Note that Champagne and Tonkens do not argue that a commanding officer would be morally responsible in virtue of them deploying the lethal autonomous weapon systems. Due to the nature of learning autonomous systems, it is still here assumed impossible or unfair to trace any harm done by the system back to a human agent. When a military commander is morally responsible on Champagne and Tonken's account, they are so solely in virtue of them having signed some contract, wherein the commander graciously and pre-emptively takes on the responsibility in exchange for the benefits of their office.

I will argue that Champagne and Tonkens' blank check approach does not provide a satisfactory solution to puzzle posed by moral responsibility gaps. Apparent moral responsibility gaps cause trouble because autonomous systems produce outputs that can have serious consequences, but for which no human agent seem morally responsible. As such, a proper answer to the puzzle posed by moral responsibility gaps must be: who is morally responsible for the outputs of an autonomous system, and why? As clarified above, Champagne and Tonkens solution to this puzzle consists of designating a volunteer to pre-emptively take responsibility in exchange for wealth or power.

However, designating a volunteer to take responsibility does not constitute actually identifying the morally responsible party. This may be made abundantly clear with the use of an example. Suppose it is opening night at the opera. The theatre is filled with guests enjoying the performance, when suddenly the old chandelier crashes down from the baroque ceiling and maims a large group of guests. After such a horrendous event, one might naturally expect people to want some answers. It might be questioned who is at fault for the chandelier tumbling down or who is morally responsible for the whole torrid affair. Imagine now that the owners of the opera do not launch an investigation into the chandelier crash. Instead, they pay a stage worker to take responsibility. As such, in response to questions about who is morally responsible for the crash, the stage worker takes the stage and announces: 'I take responsibility – I am sorry!'.

The stage worker's statement does not answer the initial question of who is morally responsible for the chandelier crashing. The chandelier could have crashed for a variety of reasons: neglect with respect to proper building maintenance, maybe someone just accidentally tripped and loosened the wrong ropes, or maybe a nefarious person caused the crash on purpose to sabotage the show. Whatever the reason, one thing is certain: the stage worker's statement does not help us clarify what actually happened or who was responsible. His paid response is, fundamentally, nothing more than stage dressing. In this way, it should now be clear how Champagne and Tonkens' theory fails. Instead of investigating who is actually morally responsible for the outputs of autonomous systems, their theory merely suggests paying someone off in wealth or power to take the fall for whatever may happen. However, a 'fall guy' is not equal to identifying the genuine morally responsible party in a situation, and as such Champagne and Tonkens fail to solve the conundrum posed by moral responsibility gaps.

Here, one might raise an objection based on the purpose of Champagne and Tonkens' paper. Their account was written in response to Sparrow (2007), who argued that the lack of a responsible party should lead to a ban on the use of LAWS in modern and future warfare. They prescribe a volunteer system in form of the 'blank check' approach, in order to argue against a ban on the development and use of LAWS. As such, in effect 'the blank check' theory is just supposed to guarantee that somebody can be held responsible when it comes to the use of autonomous systems. In other words, the theory is only supposed to be a contingency plan for dealing with moral responsibility gap problems.

While this may be true, in effect this objection simply concedes the point at issue by taking Champagne and Tonkens to be assuming that there are indeed moral responsibility gaps and acknowledging that there is still a need in situations featuring these gaps to pin the 'blame' for the consequences of the outputs of autonomous systems on somebody. Hence, they suggest that a designated volunteer can take the responsibility in such situations, allowing for the practical use of autonomous systems. The account does not solve the moral responsibility gap by showing that a given agent was actually morally responsible all along. They presume that there is no solution to the moral responsibility gap and therefore arrive at the conclusion that a volunteer must take one for the team, so to speak. In other words, Champagne and Tonkens fail to genuinely close the moral responsibility gap – they have instead just admitted that these gaps cannot be closed, that is, that there is nobody who is genuinely morally responsible for the outputs of autonomous systems, thereby triggering the need for their ad-hoc solution. This admission furthermore constitutes an assumption that the moral responsibility gap is unsolvable – an assumption that Champagne and Tonkens are not entitled to considering the lack of clear research on the area. Hence, Champagne and Tonkens fail to solve the problem that we started out with: is anybody morally responsible for the outputs of autonomous systems?

On that note, I will leave Champagne and Tonkens' account of responsibility and lethal autonomous weapon systems for now, and move on to autonomous vehicles. Returning to Hevelke and Nida-Rümelin's aforementioned paper (2014), they present two separate arguments for the claim that implementors can be held morally responsible in case of collisions featuring autonomous vehicles. In the following discussion, it should be noted that the term 'implementor' in cases featuring autonomous vehicles refers to the human agent occupying the vehicle at the given time.

In the first argument, Hevelke and Nida-Rümelin argue that the agent occupying the autonomous vehicle could be morally responsible through a duty to intervene. For autonomous vehicles that are on the market at the time of writing, the occupier is still instructed to be behind the wheel in the car, and to take control of the car if deemed necessary. Therefore, placing responsibility through a duty to intervene would be relatively practical for the use of current autonomous vehicles. For the purpose of being consistent with Hevelke and Nida-Rümelin's terminology, the term 'driver' will be used interchangeably in this section with the 'occupier of the autonomous vehicle', though 'driver' as a term must be considered a misnomer when used in relation to autonomous vehicles.

First, Hevelke and Nida-Rümelin (2014) attempt to motivate a duty to intervene on the part of the driver, such that the responsibility in the case of a collision would be grounded in the driver's lack of attention and appropriate intervention. To motivate this duty, they write: "If the introduction of autonomous vehicles reduces accidents by fifteen percent, and a duty to intervene for the 'driver' would lower the death rate by another fifteen, that would seem to create a moral obligation on drivers to be on the lookout for possible failure." (Hevelke and Nida-Rümelin 2014, 624). By subscribing to a consequentialist-style framework, Hevelke and Nida-Rümelin argue that

as long as a duty to intervene is shown to decrease collisions featuring autonomous vehicles, the driver can legitimately be held to be the morally responsible party (2014, 625).

I will argue that Hevelke and Nida-Rümelin's theory fails to refer to the moral responsibility of interest in moral responsibility gap problems. Recall that moral responsibility gap problems are concerned with the question: who is morally responsible for the specific outputs of a given autonomous system? I will argue that Hevelke and Nidä-Rumelin's account fails to answer this key question. To see how, let us start by using an example. Imagine a woman, Jane, walking next to her friend during a nice stroll. As they walk past a stranger, the friend suddenly and unprompted stabs the stranger, who later dies of their injuries. Further, assume that the friend was not suffering from any mental incapacity that would strip her of moral responsibility for the stabbing, and that Jane has a duty to intervene in situations such as this.

In light of this story, consider then the following questions:

1. Is Jane morally responsible for stabbing the stranger?
2. Is Jane morally responsible for not intervening, such that the death of stranger did not come about?

It is clear that these two questions are not identical, especially considering that they may well yield different answers. Regarding question 1, it is obviously true that Jane is not morally responsible for the actual act of stabbing the stranger. In this specific case, that responsibility is intuitively reserved for the friend who did the stabbing. Regarding question 2, however, since we are assuming that Jane had a duty to intervene, she can be morally responsible with respect to her failure of intervening.

These two questions take the following form when the case is replaced with one featuring an alleged moral responsibility gap:

1. Is agent X morally responsible for the output of the autonomous system that resulted in harm?
2. Is agent X morally responsible for not intervening in such a way that would have hindered or prevented the harm coming about?

If Hevelke and Nida-Rümelin's theory were to answer the key question of who is morally responsible for the outputs of autonomous systems, then the two questions should yield the same result: the user is responsible for the output because they should have intervened to prevent the ensuing harm but did not do so. However, as we saw from the case featuring the murderous friend, these two questions are not the same. Hevelke and Nida-Rümelin focus on autonomous cars in their papers and in such cases the distinction between the two questions still shows. The question of whether or not the driver was morally responsible for not intervening, in order to prevent the car's decision-making or output from that prima facie caused the collision, is an independent question from whether they are responsible for the decision-making or output itself. Hence, Hevelke and Nida-Rümelin's account does give us someone who might be morally responsible for the consequences of an autonomous system's output, that is someone who is responsible for the lack of intervening – but it does not give us the agent who is morally responsible for the actual output of the autonomous system itself.

Here, one might wish to raise an objection. In the murder case there is a very obvious answer to the question of who is morally responsible for the action/output that caused harm, namely the friend. However, in cases featuring autonomous systems, the answer to the question is not so easily attainable, hence the very problem of moral responsibility gaps. Basically, this leads to the same problem that was just discussed in relation to Champagne and Tonkens' (2015) account. The relevance of Hevelke and Nida-Rümelin's theory hinges on there being no one who is morally

responsible for the actual output of a given autonomous system. However, as with Champagne and Tonkens, considering the lack of research on this topic such an assumption cannot be fairly made. I conclude that, while Hevelke and Nida-Rümelin's 'duty to intervene' approach might be able to find drivers responsible for a lack of intervening, in decision-making cases featuring autonomous vehicles, they do not show the driver to be morally responsible for the actual decision or system output that causes collisions.

I now turn to the second solution to the moral responsibility gap proposed by Hevelke and Nida-Rümelin. They describe this solution as follows: "… an approach in which the person in charge of the autonomous vehicle has no duty (and possibly no way) of interfering, but [can] still be considered morally responsible for possible accidents." (Hevelke and Nida-Rümelin 2014, 626). The solution, named 'Strict Liability', suggests that a driver can be held morally responsible based on them taking the risk of using the autonomous vehicle, knowing that collisions can occur. In a footnote, Hevelke and Nida-Rümelin say about the name 'Strict Liability': "We will use this legal term to refer to a moral stance…" (2014, 626). The use of legal terms should therefore not be seen as a divergence on their part from the topic of moral responsibility; they are explicitly using legal terminology to illuminate moral questions. Indeed, the mix of legal terms with philosophical jargon is a regular occurence in Hevelke and Nida-Rümelin's paper, which routinely uses 'liability' interchangeably with 'moral responsibility'.

In just a couple of paragraphs, Hevelke and Nida-Rümelin argue that the 'Strict Liability' approach is a possible solution to the moral responsibility gap. They motivate their account as follows:

"If the user were only to be responsible for taking the risk of using the vehicle, he would therefore share this responsibility with every other person in the country who does the same.

From this perspective, they did not do something wrong in the sense of it being blameworthy, but they did participate in a practise which carries risks and costs for others and it therefore is their responsibility to shoulder that burden." (Hevelke and Nida-Rümelin 2014, 626).

As this responsibility for risk is shared amongst all users of autonomous vehicles, Hevelke and Nida-Rümelin propose a tax for the users as a solution to the moral responsibility gap. This tax is then supposed to pay for any potential accidents caused by autonomous vehicles.

I will keep my discussion of this proposed solution to moral responsibility gap problems brief. First, I wish only to point out that that this 'solution', just like the two previously presented, fails to identify who is morally responsible for the actual output of the autonomous system in question, and thereby fails to provide a genuine answer to the moral responsibility gap puzzle. By Hevelke and Nida-Rümelin's own admission, the user is only morally responsible for the risk of the output of the autonomous vehicle causing harm, and not for the output itself. Second, I will argue that Hevelke and Nida-Rümelin's application of the term 'Strict Liability' in the moral domain is unwarranted and unjustifiable.

In legal practice, strict liability is standard term just referring to the following type of offence: "If *mens rea* or negligence need not be proven in respect of one or more elements of the *actus reus* of an offence, that offence is one of strict liability" (Allen and Edwards 2021, 130). Strict liability is regularly applied in relation to vehicular traffic offences, such as speeding cases, where intention to break the speed limit (the mens rea in the given case) is irrelevant for the prosecution of the defendant.

However, Hevelke and Nida-Rümelin fail to explain how strict liability might work in the moral domain. An obvious interpretation suggests that Hevelke and Nida-Rümelin wish to paint users of autonomous cars as morally responsible in the case of collisions featuring such cars while

keeping the users' intentions about the harm irrelevant, in other words attributing moral responsibility to the users without evidence of *mens rea*, i.e. without evidence of a guilty mind. Nevertheless, it is unclear how we are supposed to get from strict liability to moral responsibility.

Imagine an agent, Linda, who exceeds the speed limit due to their car having a faulty speedometer. The faulty speedometer was not spotted by the garage that recently serviced Linda's car. Linda is legally liable for speeding, as it is a strict liability type offence, however she is intuitively not morally responsible for it, since she took all the reasonable precautions one could expect to avoid speeding. Hence, strict liability does not, in general, suffice for moral responsibility. Yet, Hevelke and Nida-Rümelin assume, in effect, without argument that in the case of users of autonomous vehicles strict liability can be substituted for moral responsibility. This is particularly strange, as they themselves appear to concede that strict liability does not imply moral responsibility, since they say in the passage quoted above that the driver "did not do something wrong in the sense of it being blameworthy". But 'doing something wrong' in the sense of being culpable – as opposed to merely doing something that causes, or risks causing, harm – is precisely what is at issue when it comes to the moral responsibility gap. Without further argument, then, the leap of ascribing moral responsibility based on legal terms is unjustified.

In this section, I have presented three arguments in favour of user/implementor responsibility, and found all three insufficient in providing a satisfying answer to the question of who is morally responsible for the output of autonomous systems. The debates on both creator and user responsibility in regard to moral responsibility must therefore be recognised to have come to a bit of a standstill, with no breakthrough or convincing arguments to be found at the time of writing.

It is tempting to think that machines are tools, and that behind their outputs, use and design, humans can always be found. An early and succinct expression of this view can be found in

the notes of Lady Lovelace on Babbage's analytical engine. She wrote: "The Analytical Engine has no pretensions to originate anything. It can do *whatever we know how to order it* to perform" (Lovelace et al. 1842, 722). This 'instrumental' view of machines has long been commonplace, when considering everything from steam-powered weaving looms to the modern computer.[19] The urge to treat and analyse autonomous systems as though they are mere tools therefore comes easily. If the autonomous system is a mere tool, then it seems that the moral responsibility for its output should be traceable to its users or creators. But, as it has been shown in these two past sections, it is far from clear if and why users or creators might be morally responsible for those outputs.

The work has therefore been cut out for this thesis. The primary overarching question must be: who is morally responsible for the output of autonomous systems? It has been shown that accounts focusing solely on creator and user responsibility while relegating autonomous systems to being mere tools of humans have yielded little of use. Therefore, in this thesis I will seek to provide a new perspective on the question. More specifically, I will seek to investigate what happens when the autonomous system itself is considered a possible candidate for moral responsibility.

[19] Instrumentalism is here used in reference to the definition by Heidegger (1977).

Chapter 2: Machine Incompatibilism

In the previous chapter, I presented the puzzle of moral responsibility gaps and showed that consideration of the creators and users of autonomous systems as morally responsible parties has yielded little result. This thesis seeks to breathe some new life into this debate by instead looking at the autonomous systems themselves and their capabilities. I will investigate what happens in moral responsibility gap cases if one considers the possibility of the autonomous system itself being morally responsible.

To start, I will in this chapter focus on a common argument against the possibility of morally responsible machines. I will compare this argument to a traditional argument for incompatibilism – the Consequence Argument – and situate the question of morally responsible learning autonomous systems within the traditional debate about whether or not determinism is compatible with moral responsibility. To this end, this chapter is divided into five sections. In §I, I will present and discuss the two traditionally-assumed types of condition for moral responsibility: an epistemic condition and a freedom-relevant condition. I will then introduce an argument from Bringsjord (2008) for the claim that machines fail to meet the freedom-relevant condition for moral responsibility. I will argue that Bringsjord's position, which I will name 'machine incompatibilism' for the purpose of this thesis, is commonly assumed in the literature surrounding machines and moral responsibility.

In §II, I argue that Bringsjords' argument for machine incompatibilism is merely a machine-focused version of the classic 'Consequence Argument' for incompatibilism. I will therefore suggest that in order to respond to Bringsjord and other machine incompatibilists, a common-sense place to start is with compatibilist responses to the original Consequence Argument.

Ultimately, if there are compatibilist responses available to the original Consequence Argument, then they ought to apply equally to Bringsjord's version of it.

In §§III-V, I present three possible argumentative moves utilised by compatibilists in responses to the Consequence Argument. In each section, a notable position or account utilising the response in question will be presented and discussed. In §III, I introduce and discuss Lewis' (1981) local miracle compatibilism. I will show how Lewis attempts to dismantle the Consequence Argument by specifically targeting its premise involving the laws of nature. In §IV, I introduce Fara's (2008) dispositional compatibilism. I will show how Fara rejects the Consequence Argument based on a rejection of the 'Transfer Principle'. I will here show how some compatibilists manage to argue for the preservation of the Principle of Alternate Possibilities (PAP), while still rejecting the conclusion of the Consequence Argument. Last but not least, in §V I present the main argument of Frankfurt (1969) and his famous 'Frankfurt-style' cases. I will here show how Frankfurt-style cases allows us to avoid the thorn of the Consequence Argument by rejecting the Principle of Alternate Possibilities (PAP).

This chapter will therefore present three different potential strategies to use as the foundation for a response to the Consequence Argument and, in extension thereof, the machine-specific version of the Consequence Argument. These strategies have laid the foundation for a vast number of compatibilist accounts of moral responsibility. The next step for this thesis will therefore be to choose one such account and attempt to extend it to cover machines as well as humans. This will be the task of the Chapter 3 and 4.

I. Against Morally Responsible Machines

In this first section, I will begin the task of investigating the possibility of learning autonomous systems being morally responsible in 'decision-making cases', as introduced in the previous chapter. To do so, I will first present two types of conditions for moral responsibility commonly assumed and discussed in philosophical literature. I will then present an argument from Bringsjord (2008) that machines of any kind will never be able to fulfil at least one of these traditionally assumed conditions.

In Chapter 1, I briefly mentioned moral responsibility as a concept referred to and used in the contemporary literature on the moral responsibility gap. Due to a lack of conceptual discussion or definition in the literature on moral responsibility gaps, moral responsibility has been left as an unfortunately vague and flimsy concept. In this chapter, I will not go into a discussion about the very nature of moral responsibility, but will instead present and discuss some of the conditions that philosophers have traditionally taken to be required for it.

To start, consider the following scenario. Imagine that you are to meet up with a friend for coffee. Unfortunately, he does not show up at the agreed time and instead leaves you to wait in the rain. Now, suppose that it turned out that he had stood you up on purpose – he was well aware of the time, and he made the decision to stand you up specifically to hurt you. If that was the case, it may be assumed that his decision to stand you up could correctly be attributed to him, and you may then be unimpressed with him as a friend and fellow moral agent. On the face of it, it may be said that your friend is morally responsible for having left you in the rain to hurt you. [20]

[20] The example and notion of moral responsibility used here is common. For other examples and treatments of the concept of moral responsibility, see Fischer (ed.) 1986, Lucas 1995 and Talbert 2019.

However, alterations to this story can be imagined such that intuitively your friend might be excused from being morally responsible. Two types of such intuitive excuses are expressed by Aristotle in his *Nicomachean Ethics* (Aristotle 1985, 1109b30-1111b5). The first excusing condition is one of ignorance. The ignorance condition captures the intuition that an agent may be excused from moral responsibility if the agent is unaware, ignorant or deceived about his action or its particular circumstances.

To see how the ignorance condition might work, we'll use the story about meeting for coffee again. Imagine that after you'd been waiting in the rain, you get a phone call and it turns out your friend is suffering from acute amnesia. Your friend does not know who he is, nor does he remember the promise to meet you for coffee. Indeed, it turns out he was completely unaware that not heading out to the coffee shop at the given time had any sort of moral effect or importance. Or it might have been revealed than instead of amnesia, your friend had been deceived. Imagine that someone wanted to see you hurt and therefore lied to your friend, convincing them that your meetup had been moved to another day. In both of these cases, your friend's ignorance seems to intuitively excuse him from being morally responsible for not meeting you for coffee and for hurting you unknowingly.[21]

The second excusing condition concerns control. This condition captures the intuition that an agent is to be excused from moral responsibility if he is not in control of his actions. To draw on the coffee meetup story yet again, imagine that it turned out that your friend had been kidnapped on their way to meeting you. Instead of meeting up with you at the agreed time, he was taken and tied to a chair helplessly somewhere. In that case, your friend's lack of control would

[21] For a general introduction to the ignorance condition, or the epistemic condition on moral responsibility, see Rudy-Hiller 2018. For further debate and discussions, see Robichaud and Wieland (eds.) 2017, Ginet 2000 and Mason 2015.

intuitively excuse him from being considered morally responsible for standing you up. After all, standing you up for coffee was not in his control.[22]

As these two types of excusing conditions tend to capture intuitions about everyday attributions of moral responsibility, any theory of moral responsibility would do well to accommodate them. Contemporary accounts of moral responsibility often reflect these intuitive excuses by focusing on the following conditions for moral responsibility:

1. The epistemic condition. When discussed in contemporary literature, this condition concerns whether the agent can be considered morally responsible considering their state of knowledge of, for example, the situation they are in or the likely effects of their action.
2. The metaphysical condition or the freedom-relevant condition. This condition concerns the agent's control over their action, and whether it is appropriate to ascribe them moral responsibility based on that.

These two types of conditions are commonly taken to be individually necessary and jointly sufficient for ascription of moral responsibility (Rudy-Hiller 2018, §0). The vagueness of these two conditions stems from the fact that each type may be commonly accepted as needed for a comprehensive account of moral responsibility, yet the specific nature and details of the two are subjects of heavy philosophical debate.[23]

[22] I will discuss the control condition, or the freedom-relevant condition, at length throughout this thesis, but for a brief introduction, see Talbert 2019 as well as McKenna and Coates 2021.

[23] For some examples of discussions surrounding the epistemic condition, see Smith 1983, Mele 2010 and Levy 2014. For further examples of the debate surrounding the freedom condition, see Frankfurt 1969, Strawson 1994 and Fischer and Ravizza 1998. Though the two conditions are usually taken to be jointly sufficient for moral responsibility, a few sceptics are lingering out there; see Zimmerman 2015 for a different view on the criteria for moral responsibility and Caruso 2021 for an overview of general scepticism about moral responsibility.

In this thesis, I will investigate the freedom-relevant conditions for ascribing moral responsibility to learning autonomous systems. In other words, I will in these chapters explore whether autonomous systems could have the necessary control of their outputs to consider them as candidates for ascription of moral responsibility. As such, even if a positive account is to be written in this thesis about autonomous systems and the freedom-relevant condition for moral responsibility, this should not be taken to be a complete account of autonomous systems and moral responsibility. After all, fulfilment of the freedom-relevant condition alone is not sufficient for attribution of moral responsibility. I will discuss this further in Chapter 5.

Here I will start by making a quick presentation of what the current literature surrounding machines says about machines being potentially morally responsible or satisfying the freedom-relevant condition for moral responsibility. As mentioned in the previous chapter, the volume of literature on machines has grown in the last decade, but is thinly spread across an incredibly wide range of sub-topics; and discussion of the idea of morally responsible machines is almost non-existent. In the *Stanford Encyclopaedia of Philosophy*, Noorman (2018) notes that the topic is sidestepped in favour of literature focusing on how to create a system that, on the surface, mimics moral behaviour:

"In the absence of any definitive arguments for or against the possibility of future computer systems being morally responsible, researchers within the field of machine ethics aim to further develop the discussion by focusing instead on creating computer system [sic] that can behave *as if* they are moral agents." (Noorman 2018, §2.2)[24]

[24] For examples of the literature on creating 'moral' machines, see Wallach and Allen 2009, McLaren 2011 and Tonkens 2012.

Discussion of morally responsible machines is more often than not relegated to throw-away comments in papers relating to the moral responsibility of creators and users of machines.[25] To demonstrate such comments' lack of philosophical detail, I will present a few examples below:

Gerdes' (2018) analysis of moral responsibility in relation to LAWS says the following about the idea of morally responsible machines with no further clarification: "contrary to humans, LAWS cannot be held morally responsible for their actions in the strong sense thereof, since this implies morality and mortality." (Gerdes 2013, 238).

In Roff's (2013) paper on moral responsibility in relation to autonomous war systems, she writes: "…for a machine to act intelligently does not make that machine a moral agent. While the machine has 'learned,' the underlying structure of the machine is still, in Kant's terms, 'determined.' And while the programmer lost the ability to control the machine, that does not change the machine's moral status." (Roff 2013, 355).

Johnson (2006) engages with the idea further than others, arguing for machines to be entities worthy of some moral considerations, but far from morally responsible themselves. On the specific topic, she writes: "Action is an exercise of freedom and freedom is what makes morality possible … Of course, this notion of human agency and action is historically rooted in the Cartesian doctrine of mechanism. The Cartesian idea is that animals, machines, and natural events are determined by natural forces; their behavior is the result of necessity. Causal explanations of the behavior of mechanistic entities and events are given in terms of laws of nature. Consequently, neither animals nor machines have the freedom or intentionality that would make them morally

[25] There are some exceptions to this rule. Considerations of morally responsible autonomous systems has previously been noted as being of current and future importance; see Hellström 2013, Gunkel 2012 and Cürüklü et al 2021.

responsible or appropriate subjects of moral appraisal. Neither the behavior of nature nor the behavior of machines is amenable to reason explanations and moral agency is not possible when a reason–explanation is not possible." (Johnson 2006, 199).

Without any further context, Gerdes (2018)'s note on autonomous systems and moral responsibility is of little use, although the quote does reflect the standard vagueness shrouding the notion of moral responsibility in the literature revolving around autonomous systems. Roff (2013) and Johnson (2006) give much clearer hints about what supposedly is the big hindrance to taking autonomous systems to be capable of morally responsible behaviour. For Roff (2013), it is allegedly the systems' determined nature which precludes the possibility of moral responsibility. I will address this claim in the next section. For Johnson (2006), it is specifically the lack of freedom that renders autonomous systems inappropriate for attribution of moral responsibility. Their mechanistic nature means their behaviour is a mere result of 'natural forces', robbing them of any freedom relevant to moral responsibility. Both Roff (2013) and Johnson (2006) are a bit vague in their dismissal of morally responsible autonomous systems, but their key objection stands clear. According to them, autonomous systems cannot fulfil the freedom-relevant condition for moral responsibility.

Bringsjord's short paper 'Ethical robots: the future can heed us' (2008) stands out by addressing the discussion surrounding machines and the control condition head-on. Bringsjord argues that a machine cannot possess autonomy or control over its actions in any of the senses traditionally connected with humans. Bringsjord considers a case featuring a robot named PERI in his lab. Using the robot's claw-like mechanism, the system has two possible outputs: it can either hold up a ball or drop it. The program determining the output is fully programmed by Bringsjord and his lab colleague. Bringsjord concludes: "It would seem that, in this experiment, whether PERI

drops or does not is clearly up to us, not him" (2008, 542). Bringsjord's first argument may be summarised in the following structure:

1. Whether PERI drops the ball or not depends solely on Peri's programming.
2. Bringsjord and his lab partner wrote PERI's programming, i.e. PERI has no control over his programming.
3. PERI therefore has no control over the output of his programming, i.e. whether he drops the ball or not.

Second, Bringsjord extends his argument to cover machines with a slightly more sophisticated decision-making process. In his second thought experiment, Bringsjord imagines PERI having an extra mechanism, which he calls a 'prover'. This extra mechanism can tell which of PERI's potential outputs is advisable at any given time. In this example, Bringsjord and his assistant do not directly write in PERI's programme whether, in a given situation, PERI drops the ball or not. Instead, PERI will drop the ball if only if the extra mechanism sends the signal to PERI that dropping the ball is advisable. Bringsjord concludes in this case: "Here again, I am mystified as to why anyone would say that PERI is free when his actions are those proved to be advisable. It is not up to him what he does: he does what the prover says to do, and humans built the prover, and set up the rules in question. Where is the autonomy?" (Bringsjord 2008, 543). Bringsjord's (2008) extended argument may be summarised as follows:

1. PERI's output (whether PERI drops the ball or not) is determined by Peri's programming and the advice mechanism.
2. Bringsjord and his lab partner wrote PERI's programming, i.e. PERI has no control over his programming.

3. Bringsjord and his lab partner created PERI's advice mechanism, i.e. PERI has no control over his advice mechanism.
4. PERI therefore has no control over the output of his programming, i.e. whether he drops the ball or not.

Bringsjord (2008) thus takes himself to have shown that robots, whether controlled directly by their programming or by an extra external mechanism, have no control over their output. While PERI in the example is a simple robot, Bringsjord's arguments may be extended to cover any robotic output that is the result of programming or an external mechanism. I will show that this is true by discussing Bringsjord's argument in relation to learning autonomous systems shortly.

Bringsjord attempts to use his PERI examples to reject the claim that machines could be 'free' or have 'autonomy' (2008, 542). While Bringsjord does not define these concepts clearly, they will be discussed further in the next section. Bringsjord ends his discussion by stating: "…the onus is clearly on anyone claiming that robots can have human-like autonomy, given that no such robot has been built, or even designed. Finally, from an engineering perspective, we have good reason to believe that the nature of robots, as artifacts programmed by us, suggests that they are human-controllable." (Bringsjord 2008, 543).

At a first glance, Bringsjord's simple PERI robot might seem like a far cry from the learning autonomous systems that are the topic of this thesis. However, Bringsjord's argument is not seemingly meant to be limited to simple systems like PERI alone. Instead, the argument is supposed to capture something particular about the nature of robots as artefacts made by humans, namely their determined nature. While learning autonomous systems are far more advanced than PERI, they are still seemingly merely determined artefacts, programmed and created by human hands. Bringsjord's (2008) argument then captures a key idea behind the classic 'instrumentalist'

notions of machines and autonomous systems. 'Instrumentalist' in this context just refers to the idea that machines are nothing more than a tool – an instrument for human agents to use (Gunkel 2017).

In this section, I will therefore first show how Bringsjord's argument covers machines in general, whose outputs are determined. Afterwards, I shall show how this basic argument extends to learning autonomous systems. For the purposes of this thesis, this first argument will be referred to as the 'basic machine argument':

1. If a machine, A, has an output that is not a result of a malfunction, then the output is determined by the machine's programming.
2. If the programming is human-made, then the machine, A, has no control over its own programming.
3. The machine, A, therefore has no control over its output.

Bringsjord's extended argument included an extra mechanism that influenced PERI's decision-making process. For learning autonomous systems, a similar extension of the basic argument can be made, such that the system's learning process is included in the argument. The expanded basic machine argument, then, takes the following form:

1. The output of a learning autonomous system (LAS) in a given situation A, if not caused by malfunction, is determined by the system's programming, the rules drawn from its learning process and its given input in situation A.
2. LAS has no control over its start programming.
3. LAS's learned rules are entailed by the system's programming and the inputs given in a series of past learning scenarios.
4. LAS has no control over its inputs in past learning scenarios.
5. LAS has no control over the rules that stems from its learning.

6. LAS has no control of the input in situation A.
7. Therefore, LAS has no control over its output in situation A.

The basic argument and the extended version thereby capture the idea that machines and even sophisticated learning systems cannot possess the type of control of their outputs which is classically connected to moral responsibility.

In this section I introduced the two key conditions for moral responsibility, the epistemic condition and the freedom-relevant condition, the latter being the condition that will be my main concern in the rest of this thesis. I showed that Roff (2013), Johnson (2006) and Bringsjord (2008) all claim, in effect, that autonomous systems cannot fulfil the freedom-relevant condition on moral responsibility. I then presented Bringsjord's argument for this claim and showed how to extend it to autonomous systems. In the rest of this chapter, I will investigate the extended basic machine argument further.

II. Incompatibilism's New Clothes

Following Bringsjord's arguments, machines – be those PERI or a LAS – lack control over their behaviour since their actions are entailed by their programming, environment, and similar factors. In other words, the machines discussed in this chapter are determined systems. By a 'determined system', I merely mean a system adhering to the following thesis:

> (D): In a determined system any present event is necessitated by previous events and laws of nature (e.g. physical laws) (Hoefer 2016, §1)

This definition requires some further clarification. First, for the purposes of this thesis, I shall understand the claim that an event or output is 'necessitated' (or, alternatively, 'determined') 'by previous events and laws of nature' to be equivalent to the following claim: that an event or output is entailed by some proposition specifying some relevant conjunction of past events plus the laws, in the spirit of van Inwagen (1983, 65). I will return to van Inwagen in a moment.

Second, in the case of 'determined systems' that are programmed, such as PERI or an LAS, their outputs are generally 'determined' at least in part by the rules of their programming. While a programming rule is directly encoded in the machine (e.g. the programme includes a specific instruction, 'in situation X, do Y' in a way that (arguably) does not happen elsewhere in nature – a falling body does not contain within itself an instruction to accelerate at a particular rate – programming rules are analogous to laws of nature in that programming rules specify a unique output for a given input, while a law of nature generally specifies a unique outcome in a given situation.

On one hand, as we have seen, the current literature is fairly sparse when it comes to machines and moral responsibility. On the other hand, the debate surrounding determined systems and moral responsibility is a vivid and thriving philosophical field with a long history. I will for the remainder of this section attempt to join these two fields by situating Bringsjord (2008) within this latter debate. I will argue that the debate surrounding moral responsibility and learning autonomous systems can be seen as an extension of the incompatibilism debate found in the philosophical literature.

To do this, I will start by drawing out the contemporary incompatibilism debate in broad strokes and present van Inwagen's classic 'Consequence Argument'. I will then argue that the extended basic machine argument is just a machine-focused version of the Consequence Argument.

I will end by arguing that this connection will justify looking to classic compatibilist accounts for a foundation for a response to Bringsjord (2008).

In everyday practice, human agents are normally assumed to be morally responsible for their actions. In relation to the freedom-relevant condition, we as people usually see each other as in control of our actions, when they are not being tied to chairs, being hypnotised and so on. However, a great deal of philosophical attention has been given to the question of whether this ordinary assumption about control would still hold if it turned out that our world (including all the agents within it) was a determined system.

Firstly, it should be noted that the truth of determinism as a thesis describing our world is heavily debated.[26] Secondly, multiple (and often assumed interrelated) questions dominate the debate on determinism. These questions all pertain to the worry that determinism could potentially rule out a list of commonly assumed human qualities: control over one's actions, free will and, of course, moral responsibility. In other words, the contemporary literature focuses on the possibility that determinism could be incompatible with these qualities of interest.

One of the most influential arguments in the debate on determinism was originally put forward by Peter van Inwagen (1983). His 'Consequence Argument' for incompatibilism has seen countless discussions and rewordings. For the purposes of this thesis, a simple control-focused version of van Inwagen's argument will do.[27] Such a version can be written as follows:

1. The past plus the laws of nature determines my present actions (i.e. determinism is true).
2. I have no control over the past.

[26] See Hoefer 2016 for general discussions surrounding determinism. See also Earman 2007 for a discussion of determinism and modern physics.

[27] For further discussions and wordings of the Consequence Argument, see Ginet 1990, Fischer and Ravizza 1998, Lamb 1977, McKenna and Coates 2021 and Vihvelin 2018.

3. I have no control over the laws of nature.
4. Therefore, I have no control over my present actions.

Compare now the Consequence Argument to the following summary of Bringsjord's basic argument extended to learning systems:

1. Input and past learning plus the rules of its programming necessitate LAS' present output (i.e. LAS is determined).
2. LAS has no control over its input and past learning
3. LAS has no control over the rules of its programming
4. Therefore, LAS has no control over its present output.

If the two arguments seem similar, this is no coincidence. The extended basic argument is merely the Consequence Argument being repurposed for the debate surrounding machines. The concepts of the past and the laws of nature in the Consequence Argument are merely replaced, in the machine specific argument, by the machine counterparts.

It is thus apparent that a pervasive assumption underlies the current literature on machines, control and responsibility, namely that machines cannot be bearers of moral responsibility because they are determined. I will call this claim, which is clearly assumed by Bringsjord (2008), Roff (2013) and Johnson (2006), 'machine incompatibilism'. A worry about contemporary machine incompatibilism can now be easily identified. Learning autonomous systems are determined. This fact alone is assumed to be incompatible with the systems' potential status as morally responsible entities. Machine incompatibilism is therefore assumed to be true without questioning. Bringsjord (2008) is merely repurposing the Consequence Argument, and machine incompatibilism, in its current state, is therefore nothing more than classic incompatibilism restricted to the specific case of machines. Thus, the literature on machines and moral responsibility

is in effect assuming classic incompatibilism to be true (or at least true of machines) without discussion or further questioning.

Although incompatibilism in the philosophical tradition has been argued for at length by many authors, its truth cannot be assumed.[28] In contrast to incompatibilism stands, of course, the compatibilist thesis – namely that determinism is compatible with free will and moral responsibility. Since machine incompatibilism hinges on traditional incompatibilism, a successful classic compatibilist response to the Consequence Argument would a fortiori undermine the argument for machine incompatibilism.

For the remainder of this chapter, I will therefore survey some classic compatibilist strategies for responding to the Consequence Argument. Due to the major influence of the Consequence Argument, any successful compatibilist account of moral responsibility would do well to utilise one of the discussed strategies – or similar – to respond to the argument. In the next chapter, I will then start the work of considering what type of compatibilist account could be extended to become a positive story of how the freedom-relevant condition for moral responsibility might apply to autonomous systems.

If a successful machine compatibilist account can be created from this – that is, an account of what it takes for a machine to satisfy the freedom-relevant condition despite its being determined by its programming and past inputs – it can be used to respond to Bringsfjord's argument and to shine a new light on the debate surrounding moral responsibility gaps. If Bringsfjord (2008), Roff (2013) and Johnson (2006) are labelled as machine incompatibilists, then this thesis can be read as a decisive attempt at machine compatibilism.

[28] For examples of other noteworthy incompatibilist arguments, see Kane 1989, Pereboom 2001 and Mele 2006.

Before, I move on to talk about the different compatibilist strategies for responding to the Consequence Argument, I wish to set the scene a bit more. More specifically, I want to delve a bit further into how the Consequence Argument is usually taken to threaten our common understanding of moral responsibility. To do so, recall the freedom-relevant condition introduced earlier. In its simplest form, this condition merely requires a morally responsible agent to be in control of their chosen actions. Traditionally, people have had the intuition that to be in control of a given action requires them to have been able to do otherwise.[29] In the free will literature, this intuition usually takes the form of the following principle.

> *The Principle of Alternate Possibilities (PAP):* The ability to do otherwise is required for free will and moral responsibility.

The concept behind PAP is simple. Recall from §I of this chapter the example of you meeting up with a friend for coffee. In the scenario where your friend was kidnapped and tied to a chair, it was said that intuitively we would not consider the friend morally responsible for standing you up. After all, one would think that he could not have done otherwise in the described scenario. Hence, the ability to do otherwise is an intuitive condition for moral responsibility.

It now should become clear how determinism threatens the concept of moral responsibility. Consider again van Inwagen's Consequence Argument: "If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us." (van

[29] See David 2020 for a breakdown of the history and development of PAP.

Inwagen 1983, 16). In other words, it seems that if the Consequence Argument is sound, then we can never do otherwise in a given situation.

The Consequence Argument notoriously makes use of a rule that van Inwagen calls 'Rule Beta' (1983, 94). In this thesis, I will refer to Rule Beta using its' other common name, 'The Transfer Principle'. In the *Stanford Encyclopedia of Philosophy*, Vihvelin writes the Transfer Principle as follows (Vihvelin 2018, **§**5):

*Transfer Principle***:** From **N***p* and **N**(*p*⊃*q*), we may infer **N***q*.

N is defined as such: "'**N***p*' abbreviates '*p* and no one has, or ever had, any choice about whether *p*.'" (van Inwagen 1989, 404). This principle is the keystone of the Consequence Argument. It is the Transfer Principle that leads the argument to the conclusion that our present acts are not up to us. The versatility of van Inwagen's argument may also be noted here. **N** can be worded to use a variety of other expressions, such as 'able to do otherwise than…', 'has a choice about whether…', 'has control of whether…', and so on.

The validity of the Transfer Principle is still debated, though the principle at first glance is very intuitive.[30] As an illustration of the principle, consider the following example. Imagine that there is a hurricane headed towards my office. Unfortunately, the presence of the hurricane entails the destruction of my office. In relation to the Transfer Principle, let *p* stand for the following sentence: 'A hurricane is going through my office'. Further, let *q* stand for 'My office is destroyed'. It is then reasonable to suggest that I have no choice about the hurricane going through

[30] For examples of this debate, see Ginet 1980, Fischer 1994, McKay and Johnson 1996 and van Inwagen 2000.

my office, nor do I have a choice about whether the hurricane's path entails the destruction of my office. As such it seems to be true that I have no choice in the destruction of my office.

Determinism according to the Consequence Argument disrupts our common view of our choices as forking paths. It seemingly removes our ability to have done otherwise in any given scenario. Metaphorically speaking, it looks like determinism necessitates our path to be straight and narrow. Whether one reads determinism as eliminating our ability to choose other paths or as removing our control over our choices, the message is clear. If the Consequence Argument is sound, then the freedom-relevant condition for moral responsibility is unattainable for determined agents.

Before discussing some compatibilist responses to the Consequence Argument, something must be said here about the relation between moral responsibility and free will. In some of the contemporary compatibilist literature, the freedom-relevant condition is seen as equivalent to free will, thereby requiring an agent to have free will in order to be morally responsible. That a link between moral responsibility and free will is usually assumed or taken for granted can be seen by the fact that compatibilism can refer to either of the following two theses:

1. Moral responsibility is compatible with determinism.
2. Free will is compatible with determinism.

The Consequence Argument is a threat to either formulation of the compatibilist thesis. As mentioned earlier, this thesis will focus on compatibilism concerning the co-existence of determinism and moral responsibility, and in particular on the question whether autonomous systems – which I assume to be determined systems – can have the kind of control or ability that allows them to satisfy the freedom-relevant condition for moral responsibility. However, I will not

assume that such control or ability would also be sufficient for free will. As such, all references in this thesis to compatibilism will refer solely to the thesis regarding moral responsibility, and this should be considered independent of a potential link to the notion of free will.

However, since the literature between moral responsibility and free will is so intertwined some of the responses to the Consequence Argument presented below will focus on the concept of freedom. As mentioned, I will not in this thesis discuss whether the mentioned types of freedom constitute free will. Instead for the purpose of this thesis, 'freedom' can for now be read as the control or ability needed for satisfying the freedom-relevant condition for moral responsibility.

In this section, I have presented a key argument from the traditional incompatibilist literature, namely the Consequence Argument. I have argued that Bringsjord's objection is merely a machine-focused version of the Consequence Argument, and labelled him and similar writers as 'machine incompatibilists'. I have thus shown that machine incompatibilism is merely an application of traditional incompatibilism to the particular case of (determined) machines. An established compatibilist response to the Consequence Argument might therefore prove equally effective in response to the machine-focused version of the argument.

In §§III-V, I will therefore provide a short overview of a selection of popular compatibilist strategies to counter the Consequence Argument. I will also comment on the different strategies' prima facie viability as the first step towards creating a machine compatibilist account. The following sections will therefore paint a picture of contemporary compatibilist responses to the Consequence Argument. However, the compatibilist literature is extensive, so the picture being painted of it here must be expected to be more impressionistic than naturalist in style.

III. Lewis and the Consequence Argument

I will here, and in §§IV-V, provide a brief overview of some well-known compatibilist rejections of the Consequence Argument. I will especially focus here on the strategies used to fuel these rejections. As the three strategies all refer to PAP, I will here focus on the 'ability to do otherwise' version of the Consequence Argument. This version can be worded as such:

> CA1: The past plus the laws of nature determines my present action, *p* (i.e. determinism is true).
>
> CA2: I do not have the ability to render the laws of nature nor the past false.
>
> CA3: Therefore, in the present I do not have the ability to render *p* false.

One is able to get from CA2 to CA3 via the Transfer Principle – the details of which were introduced in §II of this chapter. CA3 means that for any given present action of mine, I am unable to do different. In other words, CA3 says that I do not have the ability to do otherwise. As *p* could be exchanged for any given present action, this means that I never have the ability to do otherwise vis-à-vis my current actions. Recall PAP from the previous section. PAP states that the ability to do otherwise is necessary for free will and moral responsibility. This means that the Consequence Argument rules out both acting freely and moral responsibility via PAP.

In his paper 'Are we free to break the laws?', Lewis (1981) argues for a rejection of premise CA2 - that we, as agents, are unable to render the laws of nature false. I will in this section use Lewis' argument to illustrate the first possible strategy for dismantling the Consequence

Argument: rejecting the premise requiring the ability to break the laws of nature or altering the past.[31]

At a first glimpse, rejecting this premise sounds strange; however, I will here recount and provide a simple summary of Lewis' illustrative argument (1981, 122). Lewis asks us to suppose that in a determined world his hand is lying on the desk at a particular time, t1. Following the Consequence Argument, the past and the laws of nature thereby entail that Lewis did not raise his hand just prior to t1. Consider now the two following statements.

1. At t1, Lewis was able to raise his hand at a normal pace.
2. At t1, Lewis was able to raise his hand faster than the speed of light.

Looking back at the Consequence Argument, it seems that for either of these statements to have been true, Lewis would have been able to render false some proposition about the distant past, or else render false – that is, to break – a law of nature.

For the first option, it may be surely said that at t1, Lewis did not possess the ability to change the past, i.e. anything that happened at any time t before t1.[32] But if the ability to break a law is required for (1) to be true, and Lewis has that ability, then it seems we must accept that (2) is true too. Yet manifestly Lewis was *not* able to raise his hand faster than the speed of light.

[31] Though my points here are made with Lewis' 1981 account in mind, they are extendable to other local miracle compatibilist accounts. For examples, see Graham 2008 and Pendergraft 2011. The term 'local miracle compatibilism' is here used according to Fischer 1994.

[32] Some compatibilists do argue against this. Similar to Lewis' handling the laws of nature, such compatibilists argue that the power to change the past only seems strange, until replaced by a milder understanding of this 'power'. These accounts fare similarly to local miracle compatibilism in terms of the points made later in this section. For recent examples of such accounts, see Dorr 2016 and Perry 2010.

Then, according to the Consequence Argument, (1) and (2) are equally preposterous, as they both require the same thing: the ability to break the laws of nature, an ability that human agents surely lack. However, Lewis (1981) argues that (1) and (2) are not, in fact, on par: only (2) requires such a preposterous ability. Lewis instead introduces a distinction between two theses surrounding abilities (1981, 123):

Weak Thesis: I am able to do something such that, if I did it, a law would be broken.

Strong Thesis: I am able to break a law.

For the purposes of this section, 'a law' can here be understood to be a law of nature. Lewis rejects the Strong Thesis, instead arguing for the truth of the Weak Thesis. To explain how this distinction comes into play, consider statement (2) again:

2. At t1, Lewis was able to raise his hand faster than the speed of light.

Here, the action of hand-raising faster than the speed of light would itself break a law of nature, since it is a law that any part of a human cannot travel that fast. Following Lewis' distinction, the truth of (2) would require the Strong Thesis to be true. One might therefore consider statement 2 to describe a 'law-breaking event'. The important thing here is to note that the action itself would be law-breaking. On the other side, consider (1) again:

1. At t1, Lewis was able to raise his hand at a normal pace.

Lewis argues that in order for (1) to be true, only the Weak Thesis needs to be true. In order for it to be true that he was able to raise his hand at normal pace, it only needs to be true that he was able to do something (namely, raise his hand at normal pace) such that, had he done it, a law would have been broken. It does not need to be true that he was able to do something that itself

constitutes a law-breaking event. This is because, according to Lewis, had he raised his hand shortly after t1, at t2 for example, there would have been some law-breaking event or other at some point prior to t2, which would (together with the state of the Universe at that time) have entailed that Lewis raised his hand at t2. Lewis refers to this law-breaking event as a 'divergence miracle'.

The important point is that the divergence miracle is not (nor is it caused by) Lewis' hand-raising, but instead predates it. Lewis writes: 'If anything the causation would have been the other way around' (Lewis 1981, 125). By 'the other way around', is meant that the law-breaking event would have been one of the causes of the hand-raising. According to Lewis, it is therefore true to say that if he had raised his hand at t1, it would have followed some divergence miracle. His act of raising his hand would have falsified the actual conjunction of law and the past at t1, in the sense that its occurrence is inconsistent with that conjunction. But in itself, it would not have been a law-breaking action, as any 'law-breaking' or divergence would have taken place prior to t2. Again, using Lewis' distinction from earlier, (1) merely requires the Weak Thesis – and not the Strong Thesis – to be true. As Lewis puts it: 'Thus I insist that I was able to raise my hand, and I acknowledge that a law would have been broken had I done so, but I deny that I am therefore able to break a law' (Lewis 1981, 125).

It then becomes clear how Lewis refutes the Consequence Argument. Since (assuming determinism) the only reason we had for denying that Lewis had the ability to raise his hand at normal pace was that accepting that claim would in effect attribute to him the preposterous ability to break a law, and since no such ability is in fact required, the Consequence Argument fails. Specifically, the Consequence Argument fails, because it is guilty of equivocating the Strong Thesis and the Weak thesis in premise CA2. Lewis' claims that while the Strong Thesis is false, this gives

us no reason at all to think that the Weak Thesis is false; and, indeed since no reason has been given to deny the Weak Thesis, he claims it to be true.

If we read CA2 as denying the Strong Thesis, then that premise is true – the Strong Thesis is indeed false – but the argument is invalid, since the denial of the Strong Thesis is consistent with (for example) Lewis being able at t1 to raise his hand at normal pace. On the other hand, if we read CA2 as denying the Weak Thesis, then the argument is valid but unsound, since the Weak Thesis is, according to Lewis, true. Either way, the Consequence Argument fails to establish that determined agents lack the ability to do otherwise, and hence is unsuccessful in establishing that they fail to act freely.

Lewis (1981) is not alone in accusing the Consequence Argument of equivocation and calling for a distinction between different abilities or theses present in van Inwagen's argument.[33] The challenge for these responses to the Consequence Argument is therefore to successfully defend the claim that van Inwagen is indeed guilty of equivocation.

I will here briefly consider the prospects for using this type of response to the Consequence Argument as part of a machine compatibilist account – that is, an account of what it takes for a machine to satisfy the freedom-relevant condition for moral responsibility considering its determined nature. To use Lewis' distinction or a similar one as the foundation for a machine compatibilist account would require completion of at least two large tasks.

First, one would need to defend Lewis' distinction between the ability to do otherwise in the strong sense and the weak sense. Considering the responses to Lewis on these points, it would

[33] Outside of the earlier mentioned writers, see also Horgan 1985 and Vihvelin 1988

not be an insignificantly sized task.[34] The same would hold true for any previously mentioned responses to the Consequence Argument using the same overarching strategy as Lewis.

Second, Lewis uses his distinction to show that humans' weak ability to do otherwise is compatible with determinism. Human agents' weak ability to do otherwise here refer to the ability to do things such that, were they to do them, a law would be broken. However, he does not provide a story of what is required of agents to possess weak abilities; he merely asserts, in effect, that some determined systems do possess them. Nothing in Lewis' rebuttal of the Consequence Argument explains how and when these abilities can be attributed to human agents.

The situation is the same when considering machines: if we want to know whether they actually have the ability to do otherwise, Lewis's account will not (and is not designed to) help us. Lewis has at least a prima facie case for claiming that the Weak Thesis is true of normal human beings in many or most normal circumstances; after all, we do generally take ourselves to have more than one option open to us when we ponder a menu or think about where to go on holiday, so arguably the Weak Thesis has common sense on its side. However, when considering machines, it is not clear that common sense favours the claim that machines may have weak abilities. If anything, common sense favours the opposite judgement: that machines are more like the kidnap victim from our earlier case, or someone who is coerced or manipulated, and hence are unable to do otherwise in this sense. Of course, if successful, Lewis' argument shows that we cannot directly infer machines as unable to do otherwise from the mere fact that they are determined systems. Nevertheless, in the absence of an account of what it takes to have weak abilities, there are no positive grounds for claiming CA2 false in respect to cases featuring machines. The best we can do is claim that, pending further investigation, it is currently unproven.

[34] For objections to Lewis' argument, see Beebee 2003, van Inwagen 2004, Huemer 2000 and 2004.

I will not discuss whether these tasks can be accomplished here. For the purposes of this thesis – that is, creating a machine compatibilist account to use in analyses of moral responsibility gap problems – the easiest thing to do is simply to use a compatibilist account that already tells a positive story of how an agent can fulfil the freedom-relevant condition for moral responsibility. Hence, as will also become abundantly clear in §V here and in chapter 4, there are more straight-forward options available.

This section has served as an example of how one strategy for dismantling the Consequence Argument is to reject one of its premises by suggesting false equivocation between two different types of abilities, corresponding to the Strong Thesis and the Weak Thesis. By drawing a distinction between types of abilities, this kind of strategy can undermine the Consequence Argument, if successful. At the very least, considering Lewis' account has shown that the incompatibilist conclusion is not clearly forced upon us by the Consequence Argument. With the first type of strategy illustrated, we shall turn our attention towards other strategies for countering van Inwagen's stubborn argument.

IV. Fara and Dispositional Compatibilism

The second strategy to be discussed is the rejection of the Transfer Principle. I will use Fara's dispositional compatibilism as an example of the utilisation of this strategy. Fara (2008), Smith (1997) and Vihvelin (2013) are all proponents of dispositional compatibilism. I will in this section outline only Fara's account, as despite the three accounts being slightly different, they may be seen as broadly similar in spirit. I will then further go on to discuss the possibility of using something akin to Fara's response as part of a machine compatibilist account.

To start, recall first PAP, the Principle of Alternate Possibilities, as outlined in §II. For compatibilists seeking to preserve PAP, a classical approach has been to conceive the ability to do otherwise in terms of a 'simple conditional analysis of ability'.[35] Fara (2008) summarises the classical Simple Conditional Analysis of Ability as follows:

> Simple Conditional Analysis of Ability: 'An agent has the ability to A in circumstances C if and only if she would A if she were to try, in circumstances C, to A.' (Fara 2008, 850).

Though multiple different versions of the Simple Conditional Analysis of Ability have been proposed, the analysis is generally considered inadequate in contemporary philosophy.[36] It is in the wake of the inadequacy of the simple conditional analysis that Fara proposes his own dispositional analysis of ability, which I will relay in this section. Dispositional compatibilism considers how everyday objects have dispositions dictating their behaviour in a variety of circumstances. Consider as an example a fragile glass disposed to break easily if dropped, or a cup of coffee disposed to burn the roof of your mouth if ingested too soon after making it.

Similar to inanimate objects, Fara argues that humans have dispositions (2008, 833). I may be said to have a disposition to look to both ways before crossing a busy road, or I might be disposed to always opt for tea when offered a beverage. Understanding human abilities as similar to dispositions motivates Fara's dispositional analysis of ability. In his paper, Fara investigates how the abilities of human agents may have dispositional characteristics. He proposes: '… there is *a* sense of "ability", a modal one, on which one may have the ability to do something even when one

[35] For just a few classic examples, see Ayer 1954, Hume 1975 and Moore 1912. For a brief overview of the importance of PAP in philosophical history, see McKenna and Coates 2021.

[36] For discussions of the failings of the analysis, see Chisholm 1964, Holton 2009 and van Inwagen 1983.

has tried and failed to do it.' (Fara 2008, 846). Fara (2008) argues for replacing the traditional analysis with the following:

> The Dispositional Analysis of Ability: 'An agent has the ability to A in circumstances C if and only if she has the disposition to A when, in circumstances C, she tries to A.' (Fara 2008, 848).

The difference between the Dispositional Analysis and the Simple Conditional Analysis is subtle, yet important. An agent or object merely having a disposition to behave in a certain way is not a guarantee that the disposition will manifest. Suppose that the fragile glass was carefully packed in bubble wrap and then dropped on a pillow. The fragile glass' disposition to easily break when dropped may then be supposed to not manifest in that case. The lack of manifestation does not change that the glass is still fragile and still possesses the disposition to easily break. In this specific case, external factors just masked the glass' disposition. This type of case will be referred to as a masking case. As will be shown in second, masking cases can be made with human agents as well. Furthermore, Fara claims that the following relation between dispositions and abilities obtains: 'Whenever an agent is disposed to act in a certain way, she has the ability to act in a certain way; and whenever her disposition is masked, so too is her ability.' (Fara 2008, 488). Masked abilities are summarised by Fara as follows:

> Masking: 'An agent's ability to A in circumstances C is masked iff
>
> (i) The agent tries to A;
>
> (ii) circumstances C obtain;
>
> (iii) the agent retains the ability while trying to A; yet
>
> (iv) the agent does not succeed in Aing.' (Fara 2008, 848)

For an example, consider the following case. Suppose that on a court is an experienced badminton player in the middle of an intense match. The player gets ready to make a smash at a time t1. She has done that shot thousands of times before and she possesses both the will and the bodily motor control to make it. However, suppose that right over the court a motion-detecting fan has been installed. As the shuttlecock crosses court, right before the player tries to make her shot at t1, the fan is triggered and starts whirring around. The flight of the shuttlecock is slightly disturbed, and it means that the player fails to make the smash at t1. According to Fara's dispositional reading of abilities, the player in this case has the ability to make a smash throughout the scenario, and the fan is merely masking it in the specific set of circumstances.

This, then, marks the key difference between the Simple Conditional Analysis and the Dispositional Analysis. According to the Simple Conditional Analysis, the badminton player was not able to make the shot: she tried and she failed. By contrast, the Dispositional Analysis can, as shown by the badminton example, explain how the badminton player is able to make her shot, even though she failed. Hence, Fara uses the concept of human abilities' dispositional nature to explain how human agents can have the ability to do otherwise. Fara summarises it as follows: 'For I do have the requisite intrinsic property, some property of psychomotor control, such that if I were to try, say, to scratch my nose instead of taking a sip of coffee, my trying and my having the property together would together cause me to scratch my nose; and, as far as my properties are concerned, this cause is complete: no further participation by me is required for me to scratch my nose.' (Fara 2008, 861).

It should now become clear how a dispositional compatibilist like Fara can reject the conclusion of the Consequence Argument. Recall the Transfer Principle from earlier and consider it in relation to abilities:

Transfer Principle: From **N***p* and **N**(*p*⊃*q*)), we may infer **N***q*.

Fara interprets **N**α to abbreviate 'α and no one is, or ever has been, able to make it the case that not-α' (Fara 2008, 862). Consider now again the badminton case from before. Using the Transfer Principle for that specific case, *p* can stand for 'There is fan, which takes the shuttlecock off course' and *q* is replaced by 'The badminton player fails to make a smash'. According to the Transfer Principle, we can infer that the player is not, nor has ever been, able to make it the case that she does not fail to make a smash.

According to Fara's Dispositional Analysis, however, the conclusion reached is not true: the badminton player did have the ability to make a smash at the time t1 in the case described, even though she could do nothing about the fan, and nothing about the fact that the presence of the fan guaranteed that she would fail to make the smash. The fan merely masked her ability. The Transfer Principle is thereby false according to Fara's Dispositional Analysis of abilities.

Fara summarises his view as follows: 'On my view, I am, at t, disposed to do other than take a sip of coffee when I so try if, and only if, I have at t some intrinsic property in virtue of which I (generally, usually, normally) do other than take a sip of coffee when I try. Again, some intrinsic property of psychomotor control does the job. It is in virtue of this property that (generally, usually, normally) I scratch my nose when I try, I snap my fingers when I try, I hum a tune when I try, and so on. I have the disposition, and so I have the ability to act otherwise—even on the supposition that determinism is true.' (Fara 2008, 862).

Therefore, the crucial point here, in the context of the Consequence Argument, is that the Transfer Principle is false. Even if the laws and the past – neither of which I can do anything about – determine that I perform an action *A*, it simply does not follow that I cannot do anything about

whether or not I *A*. I do have the ability to do otherwise than *A*, because I have the disposition to do otherwise than *A*, if I were to try.

This then makes up the new dispositionalists' strategy: use a dispositional reading of abilities to reject the Transfer Principle, thereby dismantling the Consequence Argument as well as preserving PAP. Their rejection of the Transfer Principle serves as the second possible response to the Consequence Argument.[37]

For the remainder of this section, I will briefly discuss some potential challenges to using Fara's dispositional compatibilism, or other accounts utilising a similar response to the Consequence Argument, as the foundation for our machine compatibilist project.

The first challenge for using dispositional compatibilism would consist in arguing that autonomous systems can be attributed dispositional abilities. Showing robotic entities in general as having dispositions seems to be a simple enough task. Consider a case of a Roomba vacuuming a living room.[38] Whenever the Roomba meets a wall, it changes direction. In that way, the following disposition can be attributed to the Roomba:

> Roomba disposition 1: The Roomba is disposed to change direction, when it hits a wall.

For a dispositional machine compatibilist account to take form (and to apply to the Roomba), one would need to argue that the Roomba not only has this disposition, but also the relevant dispositional ability:

[37] Earlier compatibilist writers have similarly tried to argue for a dismissal of the Transfer Principle based on a compatiblist account of the ability to do otherwise; see Foley 1979 and Slote 1982.

[38] 'Roomba' here refers to the popular brand of robotic hoovers.

Roomba ability 1: The Roomba has the ability to change direction.

This is not a straightforward move to make. Plenty of objects that have dispositions lack abilities; the fragile glass is disposed to break easily when dropped but it lacks the *ability* to break; glasses do not have any abilities at all. For starters, on the Dispositional Analysis having an ability to *A* requires it to be true that one would have the disposition to *A* if one were to try to *A*, and glasses are incapable of trying. Are Roombas, or more sophisticated machines, capable of trying? In order to answer that question, we would have to come up with an account of the necessary and sufficient conditions for trying that could in principle apply to beings other than humans; and that is unlikely to be a straightforward task.

The second challenge pertains not specifically to the inclusion of machines – or more specifically, autonomous systems – but instead to the relevance of dispositional abilities in general when attributing moral responsibility. The challenge here consists of the worry that even if dispositional abilities are attributable to humans and some machines alike, they might not be the relevant ability for fulfilling the freedom-relevant condition for moral responsibility.

One such worry is formulated succinctly by Whittle (2010), who makes a distinction between global and local abilities. A global ability of an object is an ability that holds true in most circumstances based on intrinsic features of the object (Whittle 2010, 2-3). As an example, although I am currently sitting down, I have the global ability to walk. 'Local abilities' does not refer to those abilities that objects or agents possess most of the time – instead a local ability is an ability in a set of given circumstances. I will here follow Whittle's notation and use hyphens to indicate the specific set of circumstances, e.g. an ability-to-do-x-in-a-set-of-circumstances-C (Whittle 2010, 3).

When using Whittle's distinction, one can make the following analysis of the badminton case presented earlier. At the time t1 in the badminton case, the player has a global

ability to make the shot, i.e. the player has the ability-to-make-a-smash-in-most-cases. Nevertheless, she does not possess the local ability to make a smash, i.e. the player lacks the ability-to-make-a-smash-when-a-fan-disrupts-the-shuttlecock-at-the-time-t1. It is in this local sense that the player cannot make a smash in the original case.

I will not go into further details here, but the keen observer might note how intuitions about moral responsibility seems linked with local, not global, abilities.[39] This is a point that the incompatibilist will want to press. I might, as Fara says, scratch my nose when I try (that is, I have the general ability to scratch my nose); but if determinism is true, I am on this occasion determined not to try: I lack the specific ability to scratch-my-nose-in-exactly-these-circumstances. Hence, the incompatibilist will insists that though I have the global ability to scratch my nose, what is required for freely refraining from scratching it is the specific ability-to-scratch-it-in-these-circumstances, which – assuming determinism – I lack. I will leave it here with the following observation: a dispositional machine compatibilist account would have to show that not only can machines be ascribed global dispositional abilities, a no small task in itself, but also that possession of such abilities fulfils the freedom-relevant condition for moral responsibility.

The first two challenges have focused particularly on the prospect of using the dispositional compatibilist response to the Consequence Argument as the starting point for a machine compatibilist account, but of course dispositional compatibilists are not alone in responding to the Consequence Argument through a dismissal of the Transfer Principle. [40]

[39] See Whittle 2010 and Kittle 2015 for further discussion.

[40] The Transfer Principle in general is perhaps the most controversial aspect of the Consequence Argument. Though Fara (2008) and other dispositional compatibilists seek to dismiss it based on the truth of a compatibilist reading of abilities, other writers have rejected the Transfer Principle on mere logical grounds. For discussions of the validity of the Transfer Principle, see van Inwagen 2000, Carlson 2000 and McKay and Johnson 1996. However, as discussed in the previous section, a rejection of the Consequence Argument alone will not be enough for the purposes of this thesis.

However, a straight-out dismissal of the Transfer Principle and in turn of the Consequence Argument is alone not enough to build a machine compatibilist account on. For the purposes of machine compatibilism, one needs not just a response to the Consequence Argument, but also a comprehensive positive story of moral responsibility. This is obvious when considering the first two challenges above. Fara's account provides a clear response to the Consequence Argument, but leaves much to be desired for a full account of how dispositional abilities are attributed and their relation to moral responsibility. Another previously mentioned dispositional compatibilist, Vihvelin (2013), does provide a comprehensive compatibilist account, yet focuses on free will. As mentioned in §III, I will in this thesis seek to stand clear of the free will debate in relation to autonomous systems to avoid confusion.

In this section, I discussed rejecting the Transfer Principle as the second type of response to the Consequence Argument. Fara's (2008) dispositional compatibilism response was used as an example of a compatibilist response to the Consequence Argument that rejects the Transfer Principle. I then raised some challenges for using dispositional compatibilism, or other arguments against the Transfer Principle, as a basis for the machine compatibilist project. To end this chapter, I will discuss one more type of response to the Consequence Argument in the next section, which I will in turn use for the creation of this thesis' machine compatibilist account.

V. Frankfurt and the Rejection of PAP

We have so far looked at two different strategies for rebutting the Consequence Argument. The first was to reject the assumption that the ability to do otherwise requires the ability to break the laws of nature. The second was to reject the Transfer Principle. This leads to the third and last category of responses that I will mention in this chapter: the rejection of PAP.

Such a strategy is utilised by Frankfurt. Frankfurt (1969) famously argued that PAP was false by the use of thought experiments of the following kind, now referred to as 'Frankfurt-style cases'.

> Black and Jones (Frankfurt 1969, 835-836): Imagine a person, Jones, who has thought and deliberated on killing another man, Smith. A nefarious doctor, Black, has an interest in making sure that Jones decides to kill Smith. To ensure his preferred outcome of the situation, Black has compromised Jones through mystical powers, a neural chip or something similar. Unbeknownst to Jones, Black now has the power to prevent Jones from deciding to not kill Smith. By a stroke of fortune (for Black, not Smith), Jones decides by himself to kill Smith, and Black's mystical powers or device never comes into play.

In cases such as this, it seems that Jones cannot do anything but kill Smith, i.e. Jones lacks the ability to do otherwise. Despite this, Frankfurt claims that intuitively Jones is still morally responsible for killing Smith (1969, 836). If this intuition holds true, then PAP is false and the ability to do otherwise is not required for moral responsibility.

Frankfurt (1969) and his rejection of PAP set the tone for a large wave of compatibilist literature. Although not uncontested, the strong intuitive force behind Frankfurt-style case has led to a large-scale abandonment of the use of PAP in contemporary compatibilist accounts. One exception to this is of course the small camp of local miracle and dispositional compatibilists, who were discussed in §§III-IV. With PAP out of the spotlight, it opens the possibility for more nuanced descriptions of what is required for the fulfilment of the freedom-

relevant condition for moral responsibility.[41] I will go into much more detail about a select few of these accounts in the following chapters.

The latter half of this chapter has served as a quick overview of the main types of responses to the Consequence Argument available to compatibilists. While nowhere close to being comprehensive, some conclusions can be drawn. First, there are clear possible argumentative routes for rejecting the Consequence Argument. As things currently stand, no such route is decisive, but we can nonetheless safely conclude that incompatibilism grounded in the Consequence Argument alone cannot be merely assumed. Yet, this assumption was seen to be prevalent amongst the few authors in the moral responsibility gap literature, who claim that autonomous systems cannot be considered for moral responsibility purely due to the systems' determined nature. They have assumed a version of the Consequence Argument restricted to machines to be decisive. This is an assumption these authors are not entitled to. If one wants to argue persuasively that autonomous systems do not, or cannot, meet the freedom-relevant conditions on moral responsibility, one needs to do better than merely assuming that the (restricted version of) the Consequence Argument is sound.

Second, there is then a fairly practical challenge for someone like me, who wishes to provide an account of freedom-relevant conditions on moral responsibility applicable to machines. The challenge is that with this specific aim, standard compatibilist responses to the Consequence Argument do not, in themselves, go very far. In particular, it is unclear how to apply compatibilist responses that seek to defend the claim that deterministic human agents are, in fact, able to do otherwise to the case of machines. This is because such responses do not, themselves, generally

[41] See Haji 2002, Levy and McKenna 2009 and Russell 2002 for overviews of the contemporary post-Frankfurt compatibilist literature. I will in the following chapters discuss a select few of the accounts treated in these.

provide any insight into what it takes to have the relevant kinds of ability, although this insight is precisely what is needed in order to establish what kinds of non-human determined systems might satisfy the freedom-relevant conditions on moral responsibility.

A safer bet therefore seems to be to follow the trend of much contemporary compatibilism and accept Frankfurt's claim about Frankfurt-style cases, namely that in such cases the agent is morally responsible despite not having the ability to do otherwise. Therefore, I will use a compatibilist account that rejects PAP as the foundation for this thesis' machine compatibilist account. The next grand quest is therefore easily set. In Chapter 3 and 4, I will present a select couple of compatibilist accounts that dismiss PAP, and I will attempt to identify one that would lend itself well to the machine compatibilist project.

Chapter 3: The Price of Leaving PAP

As we embark on a new chapter, I wish to take a brief moment in order to clarify some points about the machine compatibilist account that I intend to develop in this thesis. In Chapter 1, I presented the concept of the moral responsibility gap and the problems that this gap can pose in cases featuring contemporary and near-future autonomous systems. The goal for this thesis is to investigate the possibility of autonomous systems being morally responsible, as this is an underexplored solution to moral responsibility gap problems.

In Chapter 2, it was then shown that the key objection currently standing against the notion of morally responsible machines is grounded in the assumption that no machine will ever be able to fulfil the freedom-relevant condition for moral responsibility, because machines are inherently determined in nature. Hence, the objection assumes incompatibilism between determinism and moral responsibility. In response, I will search for a viable compatibilist account that can be extended to autonomous systems and thereby create a machine compatibilist account.

The starting point for this thesis is to find a solution for moral responsibility gap problems via entertaining the idea of morally responsible autonomous systems and as such, this puts a constraint on our search for a compatibilist account for use here. For a compatibilist account to be viable – that is, viable as the foundation for a machine compatibilist account created for the purposes of this thesis – it must allow for the hypothetical possibility of current or near-future machines fulfilling the freedom-relevant condition(s) for moral responsibility. This constraint is necessary in order to develop a machine compatibilist account that can move forward the debate on moral responsibility gaps. Having clarified our overarching goal and our condition for a viable compatibilist account to use, we can now start our search.

As mentioned in Chapter 2, I will in this thesis accept the intuitive force of Frankfurt-style cases, and therefore seek a compatibilist account that does not rely on PAP. Having left the notion of PAP behind, a glaring question stands for the machine compatibilist project: what abilities or properties fulfil the freedom-relevant condition for moral responsibility? In this chapter and the following, I will discuss accounts that all seek to answer this question.

I will discuss the possibility of using three different compatibilist accounts for the machine compatibilist project. The post-PAP wave of compatibilist literature is vast, but I will here focus on three influential positive accounts of moral responsibility, which respectively point in three of the main directions for contemporary compatibilism. The accounts I will discuss are those of Wolf (1987), Strawson (1962) and Fischer and Ravizza (1998). The first two will be discussed here, while the latter will have to wait until Chapter 4.

In §I, I will introduce Wolf's (1987) 'mesh' account of the freedom-relevant condition for moral responsibility, and discuss the possibility of applying it to autonomous systems and machines in general. I will argue that neither Wolf's (1987) account nor mesh accounts in general lend themselves naturally to the machine compatibilist project due to their requirement that the mind of morally responsible agents adhere to a specific hierarchical structure.

In §II, I will present Strawson's (1962) compatibilist account of moral responsibility. I will clarify how the account focuses on reactive attitudes, and what possible scenarios can exempt agents from being morally responsible.

In §III-IV, I will raise a series of concerns for the possibility of a Strawsonian machine compatibilist account of moral responsibility. In §III, I will discuss Strawson's 'reversal thesis' and the role of retribution in relation to autonomous systems. I will argue that Strawson's use of the reversal thesis automatically precludes autonomous systems from being morally

responsible. I will then further discuss how Strawson links reactive attitudes to retributive feelings in his account. I will argue that a Strawsonian machine compatibilist would not only have to prove the appropriateness of holding reactive attitudes towards autonomous systems, but also find a way to deal with retributive feelings towards these systems.

In §IV, I will discuss the concept of self-reactive attitudes vis-à-vis autonomous systems. I will argue that a Strawsonian machine compatibilist account would require autonomous systems to have self-reactive attitudes, such as introspective feelings of guilt in relation to moral wrongdoings. I will argue that such attitudes are not possible to attribute to current and near-future technology, thereby rendering moral responsibility according to a Strawsonian account unobtainable. This chapter will therefore end with me having demonstrated multiple objections against the practical possibility of a Strawsonian machine compatibilist project.

I. Wolf and the Criterion of a Sane Deep-Self

In this section, I will introduce and discuss Wolf's (1987) 'sane deep-self' account of moral responsibility. I will discuss how Wolf dismisses the worries raised by the Consequence Argument, and consider the possibility of extending a mesh account such as Wolf's to cover autonomous systems. I will in the end argue that neither Wolf's account nor similar mesh accounts are intuitively useable for the analysis of moral responsibility for autonomous systems.

Following Frankfurt's (1969) rejection of PAP, as discussed in Chapter 2, the philosophical literary scene fostered new attempts at developing compatibilists accounts – this time without requiring the ability to do otherwise. These attempts include Frankfurt's (1971) own influential account, wherein the kind of freedom required for moral responsibility is instead

grounded in actions stemming from desires that suitably 'mesh' with aspects of the agent's psychology. As such, Frankfurt became a key proponent of a wave of 'mesh' theories, where responsible agency requires one's desire to mesh with certain structural aspects of one's psychology or mind. [42] Wolf's account is one such account that follows Frankfurt's (1971) original line of enquiry.

In the spirit of Frankfurt, Wolf's (1987) account does not rely on PAP. Wolf starts outlining her account by observing that moral responsibility is not solely dependent on our will and control of our actions. She summarises this as follows: "…the key to responsibility lies in the fact that responsible agents are those who for whom it is not just the case that their actions are within the control of their wills, but also the case that their wills are within control of their *selves* in a deeper sense" (1987, 375).

According to Wolf, this is what differentiates a morally responsible person from a mere animal or machine: a possession of a deeper self that governs our will. As such, Wolf adds a structural element to her conception of moral responsibility. Morally responsible agency is dependent on the agent's action stemming from a deeper self. The concept of a deeper self is easily explainable. Suppose that in this moment I have a desire to smoke. At the same time, it is conceivable that I do not want to have this desire. This second want is an expression of my deeper self. This structural aspect of Wolf's account makes it very recognisable as belonging to the category of ' mesh' accounts. It is worth noting, however, that Wolf's account differs most obviously from Frankfurt's in this aspect as well. While Wolf focus on one's desires 'meshing' with

[42] Other examples of mesh theories can be found in Bratman 2007, Dworkin 1970 and Velleman 2002.

a deeper self, Frankfurt requires the 'mesh' to be with one's second-order desires – in other words, one's desire to desire something. [43]

Wolf argues, however, that being in possession of a deeper self, which governs our will, is in itself not enough for moral responsibility. To support her argument, Wolf lays out the following case (1987, 379-381). Imagine a dictator named JoJo, who since birth has followed in his father's footsteps. His father preceded him on the seat of power and led the country with no regard for human lives or well-being. Having learned everything from his dad, it is no surprise that JoJo develops warped values through his life. As a ruler, he sends people to be executed or to labour camps without a second's thought or hesitation.

JoJo acts according to his desires and will. He is neither forced nor coerced. Most importantly, his actions are expressive of his deepest self. In this regard, he differs from addicts, hypnosis victims and similar cases, where the agents' actions are alienated from the agents' deepest self. Wolf argues this point: "However, we cannot say of JoJo that his self, qua agent, is not the self he wants it to be. It *is* the self he wants it to be. From the inside, he feels as integrated, free and responsible as we do" (1987, 380). JoJo the dictator does not see his careless execution of civilians as an evil. His values are so warped from his childhood learning that he is in no position to rightly identify his own actions as cruel.

If possession of a will governed by one's own deeper self is a sufficient condition for moral responsibility, then JoJo fits the criteria. Yet, Wolf hesitates to attribute to him moral responsibility for his actions. She argues: 'In light of JoJo's heritage and upbringing – both of which he was powerless to control – it is dubious at best that he should be regarded as responsible for what he does. It is unclear whether anyone with a childhood such as his could have developed

[43] See Frankfurt 1971 for further explanation of first- and second-order desires.

into anything but the twisted and perverse sort of person that he has become' (1987, 379). For Wolf, this intuition that JoJo cannot be fully morally responsible for his actions means that an agent having a deeper self that governs their will is not sufficient for moral responsibility.

Wolf addresses this worry by introducing a 'sanity' condition on the attribution of moral responsibility. Wolf describes sanity as follows: "Sanity… involves the ability to know the difference between right and wrong, and a person who, even on reflection, cannot see that having someone tortured because he failed to salute you is wrong plainly lacks the requisite ability" (1987, 382). Wolf therefore adds that to be attributed moral responsibility, the agent in question must be sane. Or in other words, the sanity condition requires one to respond appropriately to reasons about right and wrong. Hence the effect of the sanity condition can be recognised as similar to that of Fischer and Ravizza's (1998) concept of reasons-responsiveness, which I will talk about in much detail in Chapter 4.[44] The condition of sanity can explain why JoJo the dictator may not be responsible for his cruel actions, as JoJo's deepest self is not fully sane. JoJo's deepest self is based on warped values that he cannot help having, nor can he revise them – as such he is unable to act using sane reasoning – and therefore one should not hold him responsible for acting on these values.

Wolf argues that the deep-self view along with the condition of sanity is all that is needed for attributing moral responsibility. Possession of a deep self which governs your will, along with being sane, are therefore co-sufficient conditions for morally responsible agency according to Wolf (1987, 382).

Having established Wolf's account of moral responsibility, it is here worth noting how Wolf dismisses the worry of determinism, and along with it, the Consequence Argument. The

[44] See Wolf 1990 and Nelkin 2011 for a more in-depth discussion of the relation between sanity and sensitivity to moral reasons.

concept of the deep self governing an agent's will is the key for Wolf's dismissal of incompatibilist woes. She writes: "Determinism implies that the desires which govern our actions are in turn governed by something else, but that something else will, in fortunate cases, be our own deeper selves" (Wolf 1987, 377).

By using her 'sane deep self' account to attribute moral responsibility, Wolf therefore rejects worries about incompatibilism. To demonstrate, Wolf concludes: "… although we may not be *metaphysically* responsible for ourselves – for, after all, we did not create ourselves from nothing – we are morally responsible for ourselves for we are able to understand and appreciate right and wrong, and to change our characters and our actions accordingly" (Wolf 1987, 384).

Wolf denies that that the ability to do otherwise is necessary for moral responsibility, instead relying on the notion of the sane deep self. As Wolf rejects PAP, another criterion for the freedom-relevant condition for moral responsibility must be identified. On Wolf's account, she offers up the criterion of being in possession of a sane deeper self, which governs one's will, as a necessary condition for moral responsibility. On Wolfs' account, having this deeper self then gives the agent the kind of control or 'freedom' that can replace the ability to do otherwise. Hence, Wolf's account allows agents to be morally responsible even if determinism is incompatible with the ability to do otherwise. The account in this way sidesteps the Consequence Argument by rejecting PAP and replacing it with the sane deep self condition.

In this thesis, I will not make use of Wolf's 'sane deep self' view as the foundation for a machine compatibilist account. One might here recall the goals for the development of a machine compatibilist account in thesis, as they were stated in the introduction of this chapter. I will here raise two concerns about Wolf's account, showing the potential use of this account to be unappealing for the machine compatibilist project.

First, I will here raise a concern about Wolf's definition of sanity as an ability. I will argue that Wolf cannot reject PAP by referring to an agent's sane deeper self. This concern does not relate to its applicability to autonomous systems, but is instead a general concern. One might suggest the following about Wolf's account: if sanity includes the ability to understand right from wrong, then the account might imply that only insane people could ever perform a morally wrong action. Wolf herself summarises the worry: 'But, it will be objected, there is no justification in the sane deep-self view for regarding only horrendous and stomach-turning crimes as evidence of insanity in its specialized sense. If sanity is the ability cognitively and normatively to understand and appreciate the world for what it is, then *any* wrong action or belief will count as evidence of the absence of that ability.' (1987, 387). In other words, the worry might be that according to Wolf's account, even the tiniest immoral act or thought is an expression of insanity. Since all wrongdoers are labelled insane, nobody can be held morally responsible for their wrongdoings.

Wolf addresses this concern by saying: 'This brings out the need to emphasize that sanity, in the specialized sense, is defined as the *ability* cognitively and normatively to understand and appreciate the world for what it is. According to our commonsense understandings, having this ability is one thing and exercising it is another…' (Wolf 1987, 387). The reason why JoJo in the above example fails to meet the sanity condition, then, is not that he in fact fails to 'appreciate the world for what it is', but that he simply cannot.

However, I will argue that this response to the original objection endangers Wolf's account. I will show how Wolf's suggestion of sanity as an ability risks sending the account back into the same metaphysical deadlock that it originally attempted to leave behind. Wolf acknowledges that defining sanity as an ability is potentially worrisome, but still a useful philosophical move to make. She writes: 'The notion of "ability" is notoriously problematic… At

this point, then metaphysical problems might voice themselves again – but at least they will have been pushed into a narrower, and perhaps a more manageable, corner' (1987, 387). I will show that this is not the case.

Consider the following story. Imagine Jane, who is a moral agent in a determined system. Jane is in possession of a deep self and is taken by her moral community to be sane. In other words, her actions are usually expressive of her deep self and she understands the difference between right and wrong. It may be assumed that on Wolf's theory, Jane is responsible for her actions in most ordinary situations. Imagine that one evening, Jane is at home and hears the doorbell ring. Looking out of the spyhole, she sees a person in need of help. For a variety of reasons, Jane keeps the door closed and pretends to not be home.

This, then, is how the problem might manifest for Wolf. While Jane is sane in this situation, she does not exercise her 'sane' ability, as this would have meant opening the door and helping the stranger. According to Wolf's theory, Jane should still be held morally responsible in this case. However, the crux of the classic metaphysical problem posed by determinism is our supposed lack of ability to do otherwise in a given determined situation. Following the Consequence Argument, Jane therefore could not have helped the stranger - in other words, Jane could not have exercised her 'sanity' ability. Whether or not Jane exercises this ability is seemingly then not up to her, as the situation was determined. A worry might therefore be that exercise of sanity as an ability is just the ability to do otherwise under a new name. Wolf dismisses PAP on account that one just needs actions to flow from our sane deeper self in order to fulfil the freedom-relevant condition for moral responsibility. However, the 'sanity' ability can be shown to be, in essence, an ability to do otherwise. Thus, the Consequence Argument can no longer be dismissed as irrelevant on Wolf's account.

As such, there is a missing piece in Wolf's account in order for it to be a positive compatibilist account. An explanation is needed for how a determined agent can be responsible for not exercising their sanity ability. Without a clarification on how the account is compatible with determinism, the sanity ability is just as elusive as the ability to do otherwise. A machine compatibilist wishing to use Wolf's account as the foundation for an account of moral responsibility and machines would then need to address this worry about sanity as an ability. This is not an impossible project, as there are other compatibilist strategies available as shown in Chapter 2. However, it does make the potential project of a Wolf-inspired machine compatibilist account a much more complex ordeal than we wish to take on here.

I wish now to raise the second concern – this one relating to autonomous systems specifically. I will here argue that Wolf's account is not easily extendable to cover autonomous systems, thereby showing it to be incompatible with my overall goals for a potential machine compatibilist account. It seems a clear observation that a machine compatibilist account might struggle with adapting both the deep-self view and the sanity condition into applicable concepts for machine morality. The structural aspect of a mesh theory such as Wolf's requires the agent in question to not only have some type of psychology, but also a specific hierarchical psychological structure. The concept of deep self and sanity are not ones intuitively analogous to something already found in the structure of contemporary autonomous systems.

The deep self is something intuitively belonging to some human beings only. In relation to the concept of the deep self and machines, Wolf writes: "lower animals and machines, on the other hand, do not have the sorts of selves from which actions *can* be alienated, and so they do not have the sort of selves from which, in the happier cases, actions can responsibly flow" (1987, 376). The fact that Wolf's theory primarily is meant for human agents is no surprise. However,

Wolf's criterion of a sane deep self reflects this fact, making the theory's applicability to autonomous systems tricky. The sanity or mental health of machines is a potential topic of conversation, although current writings do not reflect any contemporary or even near-future examples.[45] On a completely basic level, contemporary and near-future machines of any kind do not have the kind of deeper self that would make them capable of having self-knowledge. [46] As such, any account that heavily relies on mind-heavy notions such as deeper selves or self-knowledge will struggle to be easily extendable for analysis of autonomous systems. [47]

This improbability of combining Wolf's core concept of the sane deep self with contemporary autonomous systems leaves the prospect of using her account for the machine compatibilist project rather dire. As such, these two serious concerns about the usability of Wolf's account make up the main reason why I will not use Wolf (1987) or any other mesh accounts for the purposes of this thesis.

II. Strawson and Reactive Attitudes

In this section, I will present P.F. Strawson's famous compatibilist account and show how the account is able to bypass the Consequence Argument. In the following three sections, I will present

[45] Ashrafian 2017 is a lonesome author on this topic, however as the machines discussed in his paper can suffer everything from PTSD to a religious crisis, we can safely say that he does not work with the concept of contemporary nor even near-future machines.

[46] For a general introduction to the concept of self-knowledge, see Gertler 2011 and 2021. Though the concept of self-knowledge seems to flow into the area for the epistemic condition for moral responsibility, there are those who deny that self-knowledge is primarily epistemic. For examples, see Bilgrami 2006, Burge 1996 and Moran 2001. However, I will not try to attempt to cobble these views together with Wolf for the use for machines here.

[47] This includes mesh accounts, as previously mentioned. See also Watson 1975 for another relevant example of such an account.

and discuss three potential worries for using Strawsonian compatibilism as the foundation for a machine compatibilist account. I will at the end seek to argue that Strawsonian compatibilism is not a useful starting point for a machine compatibilist account.

Strawson (1962) argues for the use of everyday concepts to fuel a compatibilist theory. In his highly influential paper, 'Freedom and Resentment' (1962), Strawson considers the importance of the reactive attitudes that humans have to each other in relation to moral responsibility. Reactive attitudes are the commonplace broadly moral attitudes and emotional responses that agents have to others' intentions and actions towards them. Examples of these attitudes include anger, resentment and gratitude.[48]

To illustrate, consider the following situation. Imagine that you're walking in town, when a stranger pushes you such that you stumble and fall forward. If you turn around afterwards and see the stranger laughing, then you might reasonably feel resentment, anger, or indignation towards the stranger, assuming that they pushed you intending to cause you harm. However, suppose that, unbeknownst to you, you were mere seconds from being hit by an oncoming tram. The stranger had only pushed you in order to save your life. As you turn around after the incident and realise the purpose of the push, you might instead feel an immense feeling of gratitude towards the stranger.

These examples illustrate reactive attitudes that are specifically reactive to actions or intentions towards the agent in question. Call these 'personal reactive attitudes'. Strawson writes of these: 'The personal reactive attitudes rest on, and reflect, an expectation of, and a demand for, the

[48] While Strawson (1962) introduces reactive attitudes as commonplace concepts, there is still great variability in how these are understood. For examples, see Bennett 1989, Darwall 2006, Deigh 2011 and Wallace 1996.

manifestation of a certain degree of goodwill or regard on the part of other human beings towards ourselves; or at least on the expectation of, and demand for, an absence of the manifestation of active ill will or indifferent disregard.' (1962, 84).

Strawson distinguishes between three types of reactive attitudes that are usually found in common society (1962, 75-76). The first type, as mentioned, are personal reactive attitudes. The second type are the generalisation of personal reactive attitudes, named 'vicarious' reactive attitudes. These are reactive attitudes that human agents feel on behalf of someone else, e.g. moral indignation or approval. Lastly, the third type are self-reactive attitudes. These relate to the moral agent's expectations of oneself in relation to other agents. Self-reactive attitudes take the form of things such as guilt, shame and feelings of moral obligation. The key to Strawson's distinctions, and to his account more generally, is that they are not mere philosophical musings, disengaged from the real world. Instead, they are meant to be merely descriptions of our actual involvement in standard human interpersonal relationships.

Next, Strawson considers why a person might be excused or exempt from the wide range of reactive attitudes. He roughly divides such considerations into two kinds, excuses and exemptions.[49] The first kind, excuses, covers cases where the agent in question lacks either knowledge or choice. Strawson describes these reasons in terms of the excusing expressions connected to them, such as "he didn't know" or "he had to do it". Note that these two excuses each mirror one of the two individually necessary and co-sufficient conditions for moral responsibility that were introduced in the previous chapter, namely the epistemic condition and the freedom-

[49] I use the terms 'excuses' and 'exemptions' to describe Strawson's distinction roughly in line with later writers such as Watson 1987 Campbell 2005.

relevant condition. With excuses, the agent might be generally considered a fair target for reactive attitudes, but just not in the specific scenario where the given excuse is in play.

The second kind, exemptions, covers cases of diminished capacity. First, this category covers hypnosis or temporary insanity, where one could say that the agent wasn't themselves. This is cause for exemption, Strawson claims: 'We shall not feel resentment against the man he is for the action done by the man he is not; or at least we shall feel less.' (1962, 78). Second, this category further covers cases where the agent in question may be morally underdeveloped or psychologically abnormal. As an illustration of how exemption from reactive attitudes would work in such cases, consider the following example. Imagine a plane high in the air, where an ordinary adult causes chaos in the cabin. He screams, yells and throws anything unfortunate enough to be near him. Further, one might imagine that he does it all merely to create discomfort and annoyance to his fellow passengers. Such an individual might be reasonably met with some indignation or resentment from the onlooking passengers. However, if the perpetrator of the chaos is not an ill-willed adult but a very young child, then while the child's behaviour is tiresome, it seems inappropriate to resent them in the same way one would the adult. As such, the child's young age is the reason for them being exempt from the full range of reactive attitudes that would befall the adult. Instead of resentment, the other passengers might think in terms of the child needing training or education. Hence, moral underdevelopment can be cause for exemption from reactive attitudes.[50]

When someone is excused or exempt from reactive attitudes, Strawson argues that we, as moral agents, take an 'objective' attitude towards the subjects instead. In the plane example, the thoughts on the child's needed training constitute the objective attitude. Strawson writes: 'To adopt

[50] While many a compatibilist account mentions children as moral agents in development, a full theory of how humans become moral agents is seldom delved into within the moral responsibility debate. For more on the topic of moral development, see Piaget 1965, Rousseau 1979, Kohlberg 1981 and 1984.

the objective attitude to another human being to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary action, of; to be managed or handled or cured or trained; perhaps simply to be avoided…' (1962, 79).

Strawson further claims that the range of emotions one can hold towards someone else, while holding a completely objective attitude towards them, may include basic emotions. However, '… it cannot include the range of reactive feelings and attitudes which belong to involvement and participation with other in inter-personal human relationships; it cannot include resentment, gratitude, forgiveness, anger, or the sort of love which two adults can sometimes be said to feel reciprocally, for each other' (1962, 79). In other words, objective attitudes towards another agent are taken when reactive attitudes towards them are unsuitable, and hence when we take them to be exempt from moral responsibility. In this way, when holding an objective attitude towards someone, we cut them off from the intimate and more important inter-personal attitudes that agents can have towards each other. In this way, they are cut off from the core of inter-personal relations within their community.

Strawson's account of reactive attitudes has now been laid out and made clear. I will now move onwards and show how Strawson uses his commonplace concepts to deny the incompatibilist thesis and in turn the Consequence Argument. First, using his description of reactive attitudes, Strawson has made commonplace observations about attitudes tied to moral responsibility. Hence, if determinism is a thesis that threatens moral responsibility, then according to Strawson's account it threatens the legitimate deployment of reactive attitudes. Determinism as a thesis therefore raises the following question for Strawson's account: if determinism is true, must

we hold wholly objective attitudes to all determined agents, that is, every agent part of a determined system?

Imagining our world to be determined, Strawson says: 'The human commitment to participation in ordinary inter-personal relationships is, I think, too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction might so change our world that, in it, there were no longer any such things as inter-personal relationships as we normally understand them; and being involved in inter-personal relationships as we normally understand precisely is being exposed to the range or reactive attitudes and feelings that is in question.' (1962, 81).

Strawson, then, like Wolf (1987), in effect rejects PAP. Instead, moral responsibility hinges on the appropriate use of reactive attitudes in common society. Strawson specifically claims that these reactive attitudes would furthermore be unaffected by the potential truth of determinism. As the passage quoted above states, Strawson claims that even if it was shown that our world and all the agents in it were determined, this mere theoretical conviction would not change how our inter-personal relationships and our use of reactive attitudes work. As such, the possible truth of determinism would not, and could not, affect the attribution of moral responsibility, since this is based on the use of reactive attitudes.

At this point, one might wish to argue against such a conclusion. Recall Strawson's own category of excuses for reactive attitudes, which prompted excusing phrases, such as "he had to do it". Imagine a man tied to a chair unable to move, while watching a child drown in a lake before him. This man can seemingly be excused from resentment, moral indignation or guilt by the fact that he had no choice or that there was no other way. This example fits Strawson's first category of exemptions from reactive attitudes. However, if the Consequence Argument is sound,

then the same excuse may be available for all our actions. “There was no other way” seems to ring true as an excuse if determinism robs moral agents of the metaphysical aspect of moral responsibility, which in turn is then an excuse from reactive attitudes on Strawson’s own account.

Strawson rejects the claim that being deterministic makes agents exempt from reactive attitudes, however. He argues:

“For it is not a consequence of any general thesis of determinism which might be true that nobody knows what he’s doing or that everybody’s behaviour is unintelligible in terms of conscious purposes or that everybody lives in a world of delusion or that nobody has a moral sense, i.e. is susceptible of self-reactive attitudes, etc. In fact no such sense of ‘determined’ as would be required for a general thesis of determinism is ever relevant to our actual suspensions of moral reactive attitudes” (1962, 87).

In this way, Strawson argues that the use of reactive attitudes is unchanged by the potential truth of determinism, thereby rejecting the incompatibilist thesis.

In the previous chapter, I argued that machine incompatibilism is fuelled by the common understanding of ‘freedom’ and machines assumed lack of this elusive ‘freedom’. Strawsonian compatibilism does dismiss concerns of incompatibilism. It does so by rejecting PAP, and instead ties moral responsibility to the everyday practices of reactive attitudes. Strawson thereby provides a full positive compatibilist account that allows for both attribution of moral responsibility as well as determinism. The influence of Strawson’s (1962) paper on the compatibilist landscape should not be underestimated. As such, it has to be considered when questions about machines and moral responsibility crop up. Nevertheless, I will in the following sections look more closely at some concerns for using Strawsonian compatibilism for the machine compatibilist project in relation to the goal of this thesis. I will therefore end this chapter by

showing why Strawsonian compatibilism cannot reasonably be used to develop a viable machine compatibilist account in line with the goals and condition set out in the introduction of this chapter.

III. On the Reversal Thesis and Retribution

I will in this section discuss two objections for a Strawsonian machine compatibilist account that are both grounded in Strawson's explanation of moral responsibility. The first relates to what is commonly referred to as the 'reversal thesis', and I will argue that Strawson's account relies on a thesis which excludes the possibility of machines becoming part of our moral community. The second relates to how Strawson links moral responsibility to retribution and punitive measures. I will argue that genuine retributive attitudes towards machines cannot be held, thereby rejecting that machines can be the targets of reactive attitudes. As such, I will argue that a Strawsonian machine compatibilist account would have little possible traction due to Strawson's definitions of moral responsibility.

Consider the following traditional interpretation of moral responsibility as expressed by Patrick Todd:

> Traditional view: A moral agent is morally responsible when they meet some independent or objective conditions for being morally responsible. It is appropriate for their community to hold them morally responsible based on these independent facts. (Todd 2016, 209).

The traditional view has been embraced in a wide variety of forms throughout compatibilist literature. The 'independent facts' are argued to be anything from facts internal to an agent's psychology, facts based on behaviour and values to, of course, facts based on an agent's ability to

do otherwise. Some examples, which have been discussed in this thesis already, include facts about an agent's dispositional ability to do otherwise (Fara 2008) or facts about an agent's use of their sane deep self (Wolf 1987).

Strawson's view differs from this traditional take on moral responsibility. Instead, on his account an agent's moral responsibility is grounded in the reactive attitudes of their moral community. So, according to his interpretation, there are no objective conditions that people must meet in order to be morally responsible, except in the purely negative sense that they must not be 'abnormal' in a way that renders them morally incapacitated. The relation between Strawson's reactive attitudes and moral responsibility is peculiar in this sense. As moral agents, we are not met with reactive attitudes because we are morally responsible. Instead, the relation is the other way around: our moral responsibility stems from the responses of our moral community.

As Coates and Tognazzini (2012) put it: 'To be morally responsible, on this account, just is to be a member of the moral community, to be someone toward whom others feel the reactive attitudes' (2012, 6). Todd (2016) refers to this as a reversal of the order of explanation of moral responsibility. I will in this section refer to it simply as 'the reversal thesis'. [51]

It might be noted that Strawson never explicitly states the reversal thesis in his paper. I shall not go into a discussion on the appropriateness of attributing the reversal thesis to Strawson's account; the reversal thesis is a common interpretation of Strawson's concept of moral responsibility in the contemporary literature, and I shall keep to this interpretation here.[52]

[51] See Ravizza 1993, Fischer 1994 and Nelkin 2011 for a common objection against the reversal thesis.

[52] For examples see Brink and Nelkin 2013, Coates and Tognazzini 2012, Kane 2005, McKenna 1998, Pereboom 2013, Todd 2016 and Watson 1987.

Independent of whether or not Strawson succeeds in explaining moral responsibility for human moral agents, I will argue that the reversal thesis raises a problem for the possibility of using his account for machines. The problem is that human agents do not currently adopt reactive attitudes to machines, nor are they seen as members of our moral community. As such, without further criteria for how machines may become part of our moral community, it is not possible to apply Strawson's account of moral responsibility to any contemporary or near-future machines.

The truth of this might seem obvious, but I will nevertheless here expand on this point. Recall first the discussion in Chapter 1 of self-driving cars and lethal autonomous weapon systems. Consider again a case in which a human life is lost due to a decision made by one of these systems. Further, imagine the reactions of the next of kin. In cases where one human agent causes the loss of life of another, the next of kin may be easily imagined to be feeling a range of reactive attitudes as described by Strawson. Anger and resentment would be potentially two of these. One might even imagine cases in which the next of kin eventually forgives the perpetrator.

However, as things currently stand it seems almost absurd to imagine the next of kin going through this range of attitudes in relation to a machine. Imagine a person who has lost their partner in a self-driving car accident. It seems grotesque to suggest that one day the person might find the strength to forgive the self-driving car.

The problem for Strawsonian machine compatibilism, then, is clear. If human agents do not feel reactive attitudes towards machines, then, trivially, machines cannot be morally responsible. Recall, from the introduction to this chapter, that I wish to create a machine compatibilist account to help analyse moral responsibility gap problems. To this end, I need to use a compatibilist account on which one can at least entertain the notion of machines fulfilling the

freedom-relevant condition for moral responsibility. Due to the consequences of the reversal thesis, a Strawsonian machine compatibilist account would not be able to do this.

Strawson's explanation of the relation between reactive attitudes and punishment provides a second reason not to pursue a Strawsonian version of machine compatibilism. He writes:

'Indignation, disapprobation, like resentment, tend to inhibit or at least limit our goodwill towards the object of these attitudes, tend to promote an at least partial or temporary withdrawal of goodwill… The holding of them does not, as the holding of objective attitudes does, involve as a part of itself viewing their object other than as a member of the moral community. The partial withdrawal of goodwill which these attitudes entail… is, rather, the consequence of *continuing* to view him as a member of the moral community; only as one who has offended against its demands. So, the preparedness to acquiesce in that infliction of suffering on the offender which is an essential part of punishment is all of a piece with this whole range of attitudes of which I have been speaking' (Strawson 1962, 90).

Strawson thus links some of his reactive attitudes, such as resentment, to retributive inclinations. More specifically, being prepared to inflict punishment on another agent is part of holding reactive attitudes towards them. Watson (1987) highlights how severe a consequence for the notion of moral responsibility this link between the two concepts bears. He writes: 'If holding one another responsible involves making moral demand, and if the making of the demand *is* the proneness to such attitudes, then scepticism about retribution is scepticism about responsibility…' (Watson 1987, 286).

It becomes immediately obvious how this creates difficulty for the concept of a Strawsonian machine compatibilist account. Recall the cases of the next of kin to a self-driving car accident victim. In that case, it already seemed dubious to imagine the next of kin holding reactive

attitudes towards the perpetrating car. It seemed strange to imagine that a person could resent a piece of machinery, such as a self-driving car. However, when we bear in mind that Strawson links reactive attitudes such as resentment to punishment, the imaginary case just becomes absurd.

Resentment as a reactive attitude, according to Strawson, includes the withdrawal of goodwill. In other words, resentment includes an acceptance of the infliction of suffering as part of the wrongdoer's punishment. As such, holding resentment towards a self-driving car or another piece of machinery is not viable for one obvious reason: trying to inflict suffering or retributive punishment onto current or near future machines is impossible. This point and the potential link between moral responsibility and retribution can already be found within the machine literature. On his discussion of the possibility of morally responsible lethal autonomous weapon systems, Sparrow writes: '…in order to be able to hold a machine morally responsible for its actions it must be possible for us to imagine punishing or rewarding it' (Sparrow 2007, 71).

Sparrow never states where his assumptions about conditions for moral responsibility stems from. Nor does he provide a tangible account of moral responsibility that could be compared to Strawson. Nonetheless, the interesting aspect of Sparrow's account in this context is his expression of a link between moral responsibility and punishment. Sparrow notes that talk of punishment in relation to near-future machine is highly unlikely to meet the retributive demands usually set out by a moral community.

He writes: 'If a grieving relative of one of the machine's victims questions whether the machine has been punished sufficiently, it will not do to point out that it is 'suffering' because there is friction in its gears as a result of not being oiled, or that it hasn't been able to log on to the web to play chess in its spare time' (Sparrow 2007, 72).

Now recall again the case of the self-driving car and the next of kin. Imagine now that instead of a self-driving car, the perpetrating piece of machinery is a lethal autonomous weapon system. Suppose it's a drone that has bombed and killed an innocent person. On a Strawsonian machine compatibilist account, one might have to imagine that the next of kin resents the drone. However, as understood by Strawson's link between reactive attitudes and retribution, a part of this resentment is the preparedness to accept the infliction of suffering on the drone, for example the aforementioned removal of the drone's chess-playing privileges.

If Strawson's concept of reactive attitudes necessitates preparedness for inflicting retributive measures, then this is a serious obstacle for the viability of a Strawsonian machine compatibilist account. After all, it seems highly unlikely that something so trivial as removing a drone's chess privileges would be able to equate the type of retribution required.

The impossibility of meaningfully punishing machines is already a topic within machine ethics. For example, in his paper on retribution and machines, John Danaher writes: 'the increase in robotisation will lead to an increase in the causal responsibility of robots for morally harmful outcomes. Since humans are naturally inclined to find someone to retributively punish when morally harmful outcomes occur, this will lead to people desiring some appropriate target of retributive blame for acts of robot harm. But since, in many cases, neither the robots nor the manufacturers/programmers will be appropriate targets of retributive blame, a 'gap' will open up. A desire for retribution will go unfulfilled' (Danaher 2016, 305). Danaher argues that because machines cannot be punished, the introduction of autonomous systems in our society will create 'retribution gaps'.

We can now link both Sparrow's (2007) and Danaher's (2016) points back to our discussion of Strawson. The problem is clear. Strawson's reactive attitudes include making moral

demands and withdrawing goodwill when a member of the moral community makes a moral transgression. Such reactive attitudes cannot meaningfully be held against machines at the time of writing. While the very idea of humans holding any reactive attitudes towards machines – now or in the near future – seems strange, it becomes nothing less than ridiculous when one also considers Strawson's link between reactive attitudes and retributive feelings. Sparrow (2007) and Danaher (2016) similarly both point out that retributive punishment of current and near-future machines is not possible. This constitutes the second problem for the viability of Strawsonian machine compatibilism raised by his conception of the nature of moral responsibility. Its link to retribution would make it an incredibly difficult account to extend to current and near-future autonomous systems.

At this point, someone inclined to pursue Strawsonian machine compatibilism might find me ungenerous in my dealing with the possibility of making meaningful retributive demands of machines. One might point out that Sparrow does consider a range of potential punishments for LAWS. As an example, he writes: 'we might administer corporeal punishment by damaging the machine in some way, or perhaps by administering electric shocks to those electrodes through which it senses damage in combat. Finally, we might institute 'capital' punishment for the most serious crimes, such as war crimes, and destroy the machines responsible for them' (Sparrow 2007, 72).

Recall again the case from earlier with the next of kin to a LAWS victim. As mentioned, if the next of kin resents (in the Strawsonian sense) the LAWS, then this includes a preparedness to accept the infliction of punishment on the drone. If one imagines the punishment as the destruction or retirement of the drone, then resentment towards the machine may sound less

strange than the initial example. One might therefore argue along these lines that holding reactive attitudes like resentment towards machines is possible, despite my efforts to argue otherwise.

However, I will here show that preparedness to accept retirement or destruction of machines is not necessarily indicative of holding a reactive attitude towards a machine. Instead, I will show that such an attitude can be equally read as the agent in question holding an objective attitude towards a given machine. To do this, consider the following scenarios.

First, imagine a woman who has lost her mother to drug addiction followed by a fatal overdose. While going through her mother's belongings, she stumbles across the mother's kit with needles and narcotics. In a moment of anger, the daughter takes the kit and burns it.

Second, imagine that a dog has been mistreated its whole life, has been raised to be vicious and to bite anything near it. One fateful day, the dog is let loose, and it bites a child, who unfortunately dies. The parents of the child demand that the dog be euthanised.

Both of these examples highlight something important. In both scenarios, the young woman and the parents can sensibly be imagined to be feeling anger towards the object/animal in question. More interestingly, the agents in the above examples exhibit something that looks like a desire for retribution, targeted at the drugs and the dog respectively.

Similar observations have been made about human's attitudes towards non-autonomous machines in a series of curious experiments detailing social relations between humans and robots. Bartneck and Hu (2008) reported on a series of experiments inspired by the famous Millgram test and demonstrated human agents accepting and willing to inflict 'punishment' on non-autonomous machines. Despite this acceptance, the human participants expressed distress towards 'harming' the machines. Rosenthal-von der Pütten et al. (2012) recorded further that humans do

have multiple emotional reactions towards non-autonomous machines, such as empathy. This was recorded both via self-assessment and in physiological reactions. Furthermore, triggering these emotional reactions were shown not to depend on the human subject having previously interacted with the machine in question.

Clearly, then, humans can and do have emotional reactions or certain attitudes towards an object or non-autonomous machine. At a first glance this suggests that it is too hasty to reject Strawsonian machine compatibilism on the grounds that machines are not part of our moral community. After all, one might argue that we are capable of having genuine reactive attitudes, such as resentment, towards machines, and hence in some cases we do regard them as part of that community.

However, I will suggest that the woman and the parents in these examples experience anger for what has happened to their loved ones, yet this does not equate to them having inter-personal reactive attitudes towards the object/animal in question. Instead, wanting the object or animal to be put out of action can instead reflect an objective attitude taken towards them. Recall that Strawson says that adopting the objective attitude towards something is to: '…see him, perhaps as an object of social policy… ; as something certainly to be taken account, perhaps precautionary account, of; to be managed… perhaps simply to be avoided…' (Strawson 1962, 79).

The explanation that the woman and the parents are holding objective attitudes towards the drugs and the dog makes much better intuitive sense as well. After all, neither the narcotics nor the dog are morally responsible for the harm their presence has caused. Neither the lifeless drugs nor the dog can be seriously considered part of our moral community in the same manner one would consider a standard human agent to be. Merely wanting to inflict or allow infliction of destruction on a thing is not sufficient to indicate a presence of an interpersonal

reactive attitudes. It can instead merely express a wish for the object to cause no more harm or for it to be avoided.

The same conclusion can be extended to cases featuring machines. Despite the fact that human test subjects do express emotional and social attitudes towards machines in a range of experiments, Nass and Moon (2000) reported that the very same human test subjects reject the notion of these systems being part of their community. They state: 'With such clear and compelling evidence of the differences between computers and people, we have not been surprised that of the thousands of adults who have been involved in our studies, not a single participant has ever said that a computer should be understood in human terms or should be treated as a person' (Nass and Moon 2000, 82).

As such, having an emotional response towards an object or machine is not equal to expression of genuine reactive attitudes towards the thing in question, nor is it an indication of seeing the thing as a member of the moral community. This tasks the Strawsonian machine compatibilist with an extra burden. They must successfully explain why one should think that agents hold personal reactive attitudes towards machines and not just objective attitudes, as was done in the drug and the dog case.

As such, two possible worries about a Strawsonian machine compatibilist account have been carved out. Strawson's use of the reversal thesis means that to qualify for moral responsibility is just to be part of a moral community without any excusing or exempting conditions obtaining. Thus, a Strawsonian machine compatibilist would have the difficult task of showing that some actual or near-future machines including autonomous systems are part of our moral community. Further, the link between reactive attitudes and retribution would leave the Strawsonian machine compatibilist with two heavy tasks. First, they would need to answer for how retribution

and punishment involving autonomous systems would hypothetically work. Second, they would need to successfully show that human agents do hold retributive attitudes towards machines and not just objective attitudes. For the machine compatibilist project in this thesis, I am looking for a positive story of how agents are morally responsible, which leaves some room for the possibility of extending that story to autonomous systems as well. This is needed in order to get the ball moving on a machine compatibilist account, which might be able to shed light on the moral responsibility gap problems. Considering the worries raised in this section, Strawson' account does not look like the compatibilist account needed for these purposes.

IV. On Self-Reactive Attitudes

In this section, I argue that Strawsonian machine compatibilism would not be able to account for the type of systems that are the focal point of this thesis. I will show that a prerequisite for moral agency on Strawson's account is the experiencing of self-reactive attitudes, which in turn is unattainable for current and near-future robotics. The objection concerns the use and importance of self-reactive attitudes in Strawson's account.

On the topic of self-reactive attitudes, Strawson writes: 'there are self-reactive associated with demands on oneself for others. And here we have to mention such phenomena as feeling bound or obliged (the 'sense of obligation'); feeling compunction: feeling guilty or remorseful; and the more complicated phenomenon of shame.' (Strawson 1962, 84-85). Note that in relation to self-reactive attitudes, 'obligation' here is not meant to encapsulate the idea of an agent being obliged based on some independent facts. Instead, the self-reactive attitude in question is 'feeling obliged', a feeling based on the agent's interpersonal relationships within their moral community. It is such feelings that I will here argue are not possible to attribute to current or near-

future autonomous systems. I will focus my discussion on the question whether the statement, 'the machine is feeling guilty' can possibly (now or in the near future) be true, by considering contemporary attempts at machine emotion within the field of robotics.

First, a brief understanding of guilt is necessary. Any philosopher of emotion must forgive me for keeping the discussion of the nature of emotions in this section rudimentary, but a simple understanding will do for the purpose of the objection. Guilt as an emotion may manifest in multiple ways. While still keeping our discussion of guilt rudimentary, it may be useful to consider what this emotion consists of. In the *Stanford Encyclopaedia of Philosophy*'s entry on emotions, Scarantino and de Sousa detail the makeup of emotions. They write:

> "A widely shared insight is that emotions have components, and that such components are jointly instantiated in prototypical episodes of emotions… we can distinguish in the complex event that is fear an *evaluative* component (e.g., appraising the bear as dangerous), a *physiological* component (e.g., increased heart rate and blood pressure), a *phenomenological* component (e.g., an unpleasant feeling), an *expressive* component (e.g., upper eyelids raised, jaw dropped open, lips stretched horizontally), a *behavioral* component (e.g., a tendency to flee), and a *mental* component (e.g., focusing attention)." (Scarantino and de Sousa 2018, §2).

Knowing that guilt as an emotion has these different components will do for now and help in our assessment of the statement: 'the machine is feeling guilty'. As such, I will start to look at attempts at artificial emotions within robotics.

Within the field of robotics, the topic of emotions has a history of being ignored. In Newell's (1982) discussion of issues in the history of AI, he remarks on the field's lack of engagement with philosophy of emotion. He writes: 'In its genesis, AI had very little involvement

with philosophy, beyond the background awareness that comes from participation in the general intellectual culture. No philosophers of mind were involved and no technical philosophical issues were dealt with' (Newell 1982, 28). For a large part of AI's history, concepts such as emotions were dismissed as being too human and program-resistant (1982, 28).

Only in recent years has attention turned to the study of emotions within the field of robotics.[53] Two separate schools of thoughts within the AI field have been dictating the goals of the emerging literature on robot emotion. I will follow the example of Duffy (2003) and distinguish between the two schools of thought as: proponents of weak AI and proponents of strong AI. In philosophy of artificial intelligence, the strong/weak AI distinction has traditionally been tied to the question: could a machine, i.e. a robot, hypothetically ever achieve consciousness?

Proponents of strong AI believe that in principle, artificial minds could be functionally identical to human minds. In philosophy of artificial intelligence, if strong AI is correct, then a machine could in principle achieve consciousness.

In contrast, proponents of weak AI hold that artificial intelligence can in principle merely *appear* to have the capabilities of a human mind – in which case machine consciousness is in principle impossible. Within philosophy more broadly, the distinction is used commonly in debates on the possibility of strong AI.[54] The philosophical interest here stems primarily from its relation to the philosophy of mind, since even the theoretical possibility of strong AI could have significant implications for philosophical theories of mind and consciousness.[55]

[53] See Haikonen 2007 and Goya-Martinez 2016 for discussion.

[54] Especially as there is seemingly little ground to reject the notion of weak AI being hypothetically possible, see Bringsfjord and Xiao 2000.

[55] Well-known work in this area includes that of writers such as Block (1979), Dennett (1978) and Kurzweil (2002).

Within the field of robotics, the strong/weak AI distinction is used slightly differently and for a different purpose. Here, 'Strong AI' and 'Weak AI' denote the aims of different research goals. Strong AI seeks to create fully artificial animals and humans. This would include emotions, phenomenal consciousness, and all other mental capacities that together are seen as necessary for the human experience. Duffy (2003) summarises it as follows: "Proponents of strong AI believe that it is possible to duplicate human intelligence in artificial systems where the brain is seen as a kind of biological machine that can be explained and duplicated in an artificial form" (2003, 178).

For a strong AI proponent, all such capacities are nothing more than information processing, functions of the brain that one day can be fully understood and replicated artificially. Goya-Martinez describes this position within the area of artificial emotions. She writes: "those that follow the cognitive approach not only want to build a machine that appears emotional, but also that *has* emotions" (Goya-Martinez 2016, 179). In this context, 'cognitive approach' may be understood as the 'strong AI approach'. In summary, for strong AI proponents in robotics, the goal is to create genuine people from wires and machine parts.

In contrast, proponents of weak AI within the field of robotics seeks to create robots that can simulate the mental life of a human being. One such proponent is Duffy, who summarises the stance as follows: 'In adopting the weak AI stance, the issue will not be whether a system is fundamentally intelligent but rather if it displays those attributes that facilitate or promote people's interpretation of the system as being intelligent' (2003, 179).

For someone adopting the weak AI stance within robotics, a successful intelligent machine might pass sophisticated versions of the Turing test and fool humans into believing they are dealing with another human being instead of a machine. As such, genuine machine consciousness and similar are not on the list of goals for the weak AI proponent. In relation to

emotions, Weizenbaum displays this school of thought perfectly. He writes: 'Even if a computer could simulate feelings of desperation and of love, is the computer then capable of being desperate and of loving? Can the computer then understand desperation and love? …the answer is "no"' (Weizenbaum 1976, 200).

Yet, weak AI proponents still have reason to investigate the topic of simulated artificial emotions. An early study of the field identifies three such reasons: "[artificial] emotion helps facilitate believable human–robot interaction… provide feedback to the user, such as indicating the robot's internal state, goals and (to an extent) intentions. Lastly, artificial emotions can act as a control mechanism, driving behavior and reflecting how the robot is affected by, and adapts to, different factors over time" (Fong et al. 2003, 151).

As mentioned, there is an important difference in the use of the strong/weak AI distinction between philosophy and practical robotics. In philosophical contexts, the weak/strong AI distinction is used primarily in the discussion of robots' potential abilities. The questions that concern philosophers traditionally are questions such as, 'could a robot theoretically ever be everything a human is apart from biology?', or 'are our mental capacities of such nature that they can be replicated through likeness of their functions only?'. The discussion relates to the theoretical definitions of mind and what these entail for the hypothetical possibility or impossibility of creating strong AI.[56]

Within robotics, most discussions are not geared towards the theoretical possibility of strong AI. Instead, in practice, weak and strong AI might be seen as goals that motivate and inspire:

[56] Philosophy of mind and philosophy of AI are both thriving, interesting areas, however I will not delve into these areas here. For an overview of philosophy of AI, see Bringsjord and Govindarajulu 2019. For discussions of functionalism - a philosophical stance often haphazardly assumed within robotics as demonstrated by this section - see Putnam 1967, Rey 1997 and Chalmers 1996.

the distinction is used as two different benchmarks for the engineering. Understanding the difference between philosophy and robotics' use of the weak/strong AI distinction is imperative. It will help explain why some of the robotics projects that will be mentioned in this section might completely ignore theoretical questions that seem foundational and in need of answering to someone used to philosophical methods of research. The weak/strong AI distinction will also further explain the variety of motivations behind the projects that will be discussed here.

Having introduced a new distinction, it is worth restating here the point of the objection that will be developed in this section. I previously mentioned that I will be arguing that Strawsonian machine compatibilism would not be able to account for learning autonomous systems. Using the weak/strong AI distinction will enable me to make a much more general statement. I will in the following paragraphs show that Strawsonian machine compatibilism cannot be used for anything less than successful strong AI.

As mentioned, the focal point of this objection is the self-reactive attitudes in Strawson's account. To start, I will therefore here discuss some different contemporary projects within the area of robot emotions. The development of emotional robots in contemporary robotics is closely connected to the development of social robots. The Director of MIT's Personal Robots Group, Cynthia Breazel, defines the category of social robots as follows:

"We argue that people will generally apply a social model when observing and interacting with autonomous robots. Autonomous robots perceive their world, make decisions on their own, and perform coordinated actions to carry out their tasks. As with living things, their behavior is a product of its internal state as well as physical laws. Augmenting such self-directed, creature-like behavior with the ability to communicate with, cooperate with, and learn from people

makes it almost impossible for one to not anthropomorphize them (i.e., attribute human or animal-like qualities). We refer to this class of autonomous robots as social robots…" (Breazel 2003, 168).

In the highest category of social robots, one finds sociable robots. These are "socially participative 'creatures' with their own internal goals and motivations". (Breazel 2003, 169). A famous example of a social robot would be Kismet created at MIT in the late 90's.[57] Kismet is able to simulate a range of facial expressions. These, together with vocalisation and head movements, allow Kismet to socially 'interact' with humans and thus to give the impression of being an emotional being. In more recent years, notable sociable robots include Nexi from MIT's Personal Robots Group and Hanson Robotics' popular creation, Sophia.[58]

Nexi is a mobile-dexterous-social (MDS) robot, whose emotive facial expressions are still some of the most impressive that have been created to date. Its makers from the Personal Robots Group summarises the emotional aspect of Nexi's features as follows: "A particularly distinguishing aspect of the MDS design is the robot's socially expressive 4 degree of freedom neck and a 17 degree of freedom face – including gaze, eyelids, eyebrows, and a jaw. The face and neck design support a wide range of emotional and dialog-based expressions" (Breazel et al. 2009, 1). In a video of its first test of expressive ability Nexi goes through a range of facial expressions, while stating: "I can tell you that I am sad, mad, confused, excited or even bored – just by moving my face" (Personal Robots Group 2008).

Now, recall the descriptions of the components of emotions from earlier. Nexi and its fellow social robots like Kismet and Sophia are all able to exhibit the expressive component of

[57] See Breazel 2002 for a full breakdown of Kismet's design.

[58] See Goertzel et al 2017 for a description of Sophia's capabilities and her use in research on human-robot relations.

certain emotions (and in their interaction with humans, arguably also the behavioural component)[59]. The projects surrounding these social robots are indicative of a large part of the practical work on creating emotional AI, as its emotive focus is restricted to expression and behaviour.

This focus is not only prevalent within the practical projects of robotics but can also be seen in the field's overlap with neuroscience and computer science. This is clearly demonstrated in Arbib and Fellous' (2014) analysis of the possibility of emotional robots from a neurological standpoint. They start by making the following clarification: "we have no criterion for saying that a robot has 'feelings', we will seek here to understand emotions in their functional context … We analyse emotion in two main senses:

(1) Emotional expression for communication and social coordination.

(2) Emotion for organisation of behaviour" (Arbib and Fellous 2004, 554).

The expressive component of an emotion is not always considered the primary interest in the non-philosophical literature on robot emotions. As an example, Haikonen (2007) puts the main focus on the behavioural aspect of emotions:

"The system reactions theory of emotions proposes that combinations of system reactions lead to dynamic machine behaviour that corresponds to human emotions.... Emotional system reactions manifest themselves as typical behaviour. Curiosity would appear as the attention fixation on novel stimuli and potentially as approaching the cause of the stimuli with explorative actions. Fear would appear as the avoidance and fleeing of the fear-causing stimuli. Desire-related emotions like love and affection would involve seeking the closeness to the object of the emotion

[59] See Breazel et al 2009 as an example on the work on Nexi and its behavioural mechanisms in relation to manipulation of human test subjects.

and complying with its needs. This emotion would be useful for servant robots" (Haikonen 2007, 155).

Having highlighted practical robotics' focus on the expressive and behavioural aspects of emotions, the challenge for a potential Strawsonian machine compatibilist account starts to show. If Strawsonian machine compatibilism is to be relevant for anything less than successful strong AI, then the account would have to show that having the behavioural and expressive components of emotion is sufficient for attributing to a machine the self-reactive attitudes relevant for moral responsibility. I will consider two different ways the Strawsonian machine compatibilist might answer this task, though I will end up dismissing them both. In spite of modern robotics' focus on the purely expressive and behavioural aspects of emotions, I will argue that these aspects alone are insufficient for the attribution of the self-reactive attitudes that are found in Strawson's (1962) account. In summary, I will argue that nothing less than a successful strong AI could even be in consideration for moral responsibility on a Strawsonian machine compatibilist account.

The first tempting solution for the Strawsonian machine compatibilist would be to look towards traditional behaviourism for help. One might turn to the works of Ryle (1949), Skinner (1974) and other behaviourist heavyweights for inspiration. At a first glance, traditional behaviourism may look promising. Using a behaviourist approach, any mental term for emotions or self-reactive attitudes can be replaced by behavioural terms. From the behaviour of a machine one need not *infer* the existence of distinct mental states that produce that behaviour; rather, ascribing a mental state to a machine would be merely to ascribe it a range of behavioural dispositions. We can therefore see how the Strawsonian machine compatibilist could in principle make use of a behaviourist approach to explain how a weak-AI machine could be attributed self-reactive attitudes. Unfortunately, however, in contemporary philosophy, behaviourist accounts have long lost their

bite.[60] A successful defence of Strawsonian machine compatibilism that depends on resurrecting behaviourism is not one that is worth pursuing.

Instead, I shall focus on the Strawsonian machine compatibilist's second option for explaining how weak AI could have self-reactive attitudes. The second option, which will be discussed here, would see Strawsonian machine compatibilism finding help in Mark Coeckelbergh's account of robot responsibility. In his paper on robot emotions and moral responsibility, Coeckelbergh (2010) proposes that through imitation of behavioural and expressive cues, moral robots could potentially be created. In other words, Coeckelbergh suggests that having the appearance of being a moral agent is sufficient for being appropriately considered to be a moral agent.

Coeckelbergh describes his own account as having a phenomenological character. It is worded briefly, as is currently the standard within the area of machine ethics; however I will here suggest that it is, in effect, a simple behaviourist account of moral status. Coeckelbergh does not suggest a behaviourist reading of emotions, but instead just of moral status, and hence it stands apart from traditional behaviourist accounts. Coeckelbergh argues that although there are multiple components to emotions, it is only the aspects relating to expression and behaviour upon which moral status is attributed in common social-emotional practice. Therefore, according to Coeckelbergh, the successful appearance of emotions would be enough for attribution of moral status including attribution of self-reactive attitudes. In other words, a system featuring a weak AI could qualify for status as a moral agent.

[60] This is particularly exemplified by the works of writers such as Chomsky 1959 and 1971, Chisholm 1957 and Putnam 1965.

Coeckelbergh writes: “they (the designer) might create robots that learn to produce the appearance of being fully moral, including the appearance of emotions-as-cognition and emotions-as-feeling. Such robots would appear to have beliefs and the ability to judge… They would, indeed, appear human. But whatever we conclude for robots, the other important conclusion of this discussion concerns human morality: to the extent that human morality depends on emotions – both in its conditions (having the capacity) and in exercising these capacities – it does not require mental states but only the appearance of such” (2010, 239).

If Coeckelbergh’s account is right in its description of the relation between the ‘appearance’ of emotions – that is, their simulation in expression and behaviour – and the ascription of moral status, then the account could be used to support Strawsonian machine compatibilism in the following way. If humans are ascribed self-reactive attitudes based solely on their appearance and behaviour, then the same can be done with machines. According to Coeckelbergh’s account, actually having the relevant mental states, ‘emotions-as-cognition and emotions-as-feeling’, is irrelevant to the attribution of moral status. Instead, the mere ‘appearance’, – that is, simulation – of these states is enough and therefore machines could be ascribed the relevant self-reactive attitudes based on their expression and behaviour only.

I will argue that Coeckelbergh’s account does not work, and hence cannot be used in the support of Strawsonian machine compatibilism. To start, recall Strawson’s definition of the self-reactive attitudes. Specifically, he writes: “there are self-reactive attitudes associated with demands on oneself for others. And here we have to mention such phenomena as feeling bound or obliged (the ‘sense of obligation’); feeling compunction; feeling guilty or remorseful or at least responsible; and the more complicated phenomenon of shame” (1962, 84-85). I will argue that these self-reactive attitudes are defined primarily by their phenomenological aspect, without which machines

cannot be attributed these attitudes. By their phenomenological aspect, I here refer to the internal experienced aspect of these self-reactive attitudes or the emotion-as-feeling part of them. Hence, I will argue that Coeckelbergh's account cannot successfully be used to ascribe self-reactive attitudes to systems using a weak AI.

Strawson focuses on the phenomenological aspect of emotions when discussing self-reactive attitudes. On Coeckelbergh's account, it would not matter if machines did not experience self-reactive attitudes, as long as they *appeared* to an outside observer to experience them. As mentioned, Coeckelbergh assumes that knowledge of other human agents' moral status stems from observation of their expressions and behaviour in a given situation: "Our theories of emotion and moral agency might assume that emotions require mental states, but in social-emotional practice we rely on how other humans appear to us. Similarly, for our emotional interaction with robots, it might also be sufficient to rely upon how robots appear to us" (2010, 238).

Coeckelbergh is thus claiming that our 'social-emotional practise', our social and emotional engagement with other humans, relies solely on how others appear to us and thus could be extended to cover machines capable of simulating human behaviour. Hence, attribution of emotion and moral status, since it is based on mere appearance or simulation, can be applied to machines. I will reject Coeckelbergh's argument by showing that the observation and consequent attribution of mental states to humans and machines differ in a critical way.

Consider the question of how one may have knowledge of another person's mental states. This is also more commonly known as the problem of other minds. The core question is: how

can I know that other human beings have a mental life akin to my own?[61] While people may share tales of their inner life and emotions, behave in certain ways, and so on, we have no direct access to other people's mental life. Coeckelbergh provides a swift answer to this problem. He writes: "As a rule, we do not demand proof that the other person has mental states or that they are conscious; instead, we interpret the other's appearance and behaviour as an emotion" (2010, 238). In short, according to Coeckelbergh (2010), our attribution of emotions to other people is based on our immediate interpretation of their behaviour; and our treatment of other people as emotional and moral beings is also based on this immediate interpretation.

Suppose we here word a similar conundrum for robot 'minds'. The main question then becomes: how can I know that a robot has a mental life akin to my own? Coeckelbergh's answer is again simple: attribution of mental states to robots may happen exactly in the same manner as it does for other humans, namely through the immediate interpretation of their appearance alone.

Yet, I will argue that there is a stark epistemological difference between these two problems, and thereby I will show that Coeckelbergh makes a false equivocation between them. To do so, I will first for the sake of argument grant the assumption that Coeckelbergh is right in his description of social-emotional practices. In other words, I will not dispute the assumption that our ascriptions of emotions to other beings are based on their immediate appearance to us.

Even granting this assumption, however, what is crucial is that our interpretations of others' appearances as emotions can be revised or even resisted in light of relevant information.

[61] The problem of other minds is of course an influential traditional philosophical problem, however this simplified summary will serve for the purpose of this thesis. For further discussion of this area, see Davidson 1991, Hyslop 1995 and Nozick 1981.

Coeckelbergh's idea that humans interpret each other's appearances as emotions is supposedly based on ordinary social-emotional practices. Granting that, consider now the following case.

When I was around ten years old, I once found my favourite book covered in newly-added crayon drawings. My younger brother soon after admitted to the grave misdeed. Based on his expression and appearance, I thought him to be truly apologetic. I believed him to be feeling shame and regret. So we fought, as siblings are known to do sometimes. I held him fully responsible for ruining my book. Later, I learned that my brother had been completely innocent in the whole ordeal. He had merely been covering for our much younger sister. Of course, I realised then that my brother had neither felt shame nor guilt, when he apologised to me. He was merely feeling protective of our sister and had therefore chosen to shield her from my childish anger.

Now, consider a similar case, where one from the start knows the emotions on display are not real. Suppose you go see a play. Imagine that you see a man on stage confessing a murder. He explains that following his misdeed, he has been plagued by guilt and paranoia. Everything about his appearance supports his story. The man's behaviour is nervous and twitchy, and his face looks haunted. Following Coeckelbergh's account, one would naturally interpret this man as feeling guilty upon observing his appearance and behaviour.

However, we know from the start that the man in question is really an actor, practising his role as Macbeth. The 'murder victim' is in fact the fictional King Duncan. As such, we might engage with the fiction and think Macbeth is feeling guilty. However, knowing that in reality it is just a play, we rescind from interpreting what we see as the actor himself feeling guilt. After all, the actor himself does not feel guilt and shame about an event that is entirely fictional. He merely appears that way.

It is easy to imagine a whole range of similar cases. These can include everything from people lying about the sincerity of their feelings to creative stories featuring behaviour manipulation. No matter the scenario, the point is the same. Even if we immediately interpret others' behaviour and appearance as emotional, our interpretations are not infallible and may be revised or even rescinded in light of relevant information. More importantly, such revisions and recensions are commonplace in the social-emotional practices that Coeckelbergh refers to.

Recall that Coeckelbergh says: "we do not demand proof that the other person has mental states or that they are conscious; instead, we interpret the other's appearance and behaviour as an emotion" (2010, 238). My point here is that while Coeckelbergh may be right about this – I naturally interpreted my brother's appearance and behaviour as manifestations of guilt without demanding proof that he was not merely pretending – it does not follow that our natural, initial interpretation of emotional appearances are not open for reconsideration. On the contrary, if knowledge is available that contradicts our initial interpretation of another person's appearance, we revise our judgement. Moreover, if we have the relevant knowledge in advance, as we do in the case of watching a play, we refrain from making the interpretation in the first place.

Bearing this in mind, consider the problem of robot minds. The 'robot minds' of interest here are systems featuring a weak AI. As an example, I will here use the sociable robot Nexi, who was introduced earlier in this section. On Coeckelbergh's account, one may attribute emotions to her based on our interpretation of her expressions and behaviour, just as we do with our fellow humans. However, further information about Nexi's emotional states is available to us. Specifically, it is known that Nexi has no phenomenology or consciousness. We know, for example, that Nexi does not *feel* sad, even though she *looks* sad. By definition, the same would be true for any system sporting a weak AI that is capable of having the appearance of emotions.

In the case featuring the actor, relevant information meant that we rescinded the initial interpretation of the actor's appearance. The same here must be done when considering weak AIs, such as Nexi. Knowing that Nexi does not have the phenomenological component of being sad results in a recension of our initial interpretation. Just like the actor was found not to be feeling guilty, Nexi must be understood as not being sad. Appearances deceive, and our interpretations of them can be revised or rescinded to accommodate this fact. Considering the information available about the inner life (or rather, lack thereof) of systems featuring weak AI that can appear to exhibit emotional behaviour, this must result in a rescinding of our otherwise initial interpretations of their appearances.

It should now be clear how the problem of other minds and the problem of robot minds differ. As a starting point for the problem of other minds, we have no intimate knowledge of others' mental states. It is precisely this lack of knowledge which Coeckelbergh uses to motivate his claim that humans immediately interpret others' appearances as emotions. However, the same is not the case when considering the problem of robot minds. With systems featuring weak AI, further relevant information about the status of their mental life is already available. The epistemological starting points for enquiries in the two problems therefore differ. Thereby, Coeckelbergh makes a false equivocation when applying the same answer to both problems.

Coeckelbergh's theory of interpretation of appearances as emotions therefore cannot be extended to cover current sociable robots, or machines featuring weak AI more generally. The phenomenological experience of emotions and self-reactive attitudes like guilt cannot legitimately be ascribed to machines, unlike humans, based on appearance alone. As such, Coeckelbergh's account cannot be used by Strawsonian machine compatibilists to explain self-reactive attitudes in systems featuring weak AIs.

I have now briefly shown two ways a Strawsonian machine compatibilist may attempt to account for self-reactive attitudes in weak AI. Both ways have also been shown to miss their mark. Then, if a machine does not have self-reactive attitudes, then it is missing a key quality of an agent with moral responsibility as described in Strawson's account. In short, any machine without self-reactive attitudes cannot be considered a contender for moral responsibility on a Strawsonian machine compatibilist account. This automatically rules out the sociable robots that have been discussed in this section, as well as any contemporary relevant robotics projects, and indeed systems using weak AI in general.[62]

This, of course, leaves strong AI as the sole potential contender for providing usable subjects for Strawsonian machine compatibilism. I will not here give any further consideration to the prospects of a version of Strawsonian machine compatibilism that would apply only to systems featuring strong AI, as the machines that are of interest to this thesis do not belong in this category. In fact, they are nowhere near, as contemporary technology is far away from achieving strong AI. Strawsonian machine compatibilism would then only be of potential interest if considering – at best – far future technology. Since my aim is to develop an account of moral responsibility that could, at least in principle, apply to current or near-future machines in order to address the moral responsibility gap, an account that can only be entertained in relation to the science-fiction dreams of strong AI is not suitable for the purposes of this thesis.

In this chapter, I have presented and discussed both Wolf's and Strawson's popular accounts of moral responsibility. Both accounts are heavyweights on the compatibilist scene, and I have discussed why these two accounts, despite their philosophical importance, do not lend

[62] The same objection may be raised towards systems simulating emotional decision-making, see Arkin 2009, Arkin et al. 2012 and Salmeron 2015.

themselves well to the machine compatibilist project. To facilitate a conversation about contemporary and near-future autonomous systems fulfilling the freedom-relevant condition for moral responsibility, involving concepts such as deeper selves, reactive attitudes or emotions is practically not conducive to the project. In Chapter 4, I will present an account that rejects PAP, while also not involving any of these troublesome concepts. This will then be the foundation for the machine compatibilist account that I will ultimately develop in Chapter 5 and 6.

Chapter 4: Reasons-Responsive Compatibilism

In this chapter, I will present the reasons-responsive compatibilist account of Fischer and Ravizza (1998), which I will use as the foundation for the machine compatibilist account developed in this thesis. By the end of this chapter, I will have gone through all the key elements of Fischer and Ravizza's account, and I will thus have laid the groundwork for investigating the possibility of expanding the account to cover autonomous systems.

This chapter will be divided into four sections. In §I, I will focus on Fischer and Ravizza's notion of control. I will first introduce Fischer and Ravizza's 'indirect' and 'direct' challenges to compatibilism, and how these relate to the Consequence Argument. I will then present their notions of regulative control and guidance control, and show the roles of these types of control in relation to the challenges.

In §II, I will introduce the first element of Fischer and Ravizza's notion of guidance control, namely reasons-responsiveness. I will explain how mechanisms may be attributed different levels of reasons-responsiveness based on their receptiveness and reactivity to reasons. §III builds on this explanation and introduces the notion of 'moderate reasons-responsiveness', the level of reasons-responsiveness that Fischer and Ravizza argue to be sufficient for guidance control.

In §IV, I will present and discuss the second element of guidance control, namely mechanism ownership. I will introduce Fischer and Ravizza's conditions for mechanism ownership and show how these are supposed to deal with worrisome cases featuring manipulated agents. I will in the end argue that Fischer and Ravizza's account does not adequately respond to incompatibilist manipulation worries, thereby leaving this aspect of Fischer and Ravizza's account open for further discussion later in thesis. In the end, I will therefore have thoroughly gone through Fischer and

Ravizza's account in preparation for using it as the foundation for a machine compatibilist account in the following chapters.

I. The Importance of Control

Fischer and Ravizza (hereafter shortened to F&R) present a comprehensive account of moral responsibility in their book 'Responsibility and Control' (1998). Before we can talk about the application of this account to autonomous systems, a proper introduction to F&R's account must be made. As such, in this section, I will start setting the scene.

The focus of F&R's account is on the metaphysical or freedom-relevant condition for moral responsibility, which they summarise as follows: '…one could say that the freedom-relevant condition specifies that the agent must *control* his behaviour in a suitable sense in order to be morally responsible for it' (1998, 13). Recall from Chapter 2 that machine incompatibilists are primarily concerned with the idea that autonomous systems or machines are incapable of possessing the type of freedom or control that is classically linked with moral responsibility. F&R's account directly addresses this key requirement on moral responsibility by developing an account of control.

F&R voice their account as a response to two types of challenges from determinism, which they call the 'Indirect Challenges' and the 'Direct Challenges': "The Indirect Challenges contend that casual determinism rules out control and thus also moral responsibility. The Direct Challenges do not proceed via the intermediary notion of control; they argue that casual determinism rules out moral responsibility (but not in virtue of ruling out control)" (1998, 17).

The Indirect Challenges can be summarised as the various versions of the Consequence Argument, some of which were presented in Chapter 2. These challenges include a

number of variations that have not been discussed so far.[63] For the purpose of discussing F&R's account of moral responsibility, from now on I shall focus on the control version of van Inwagen's Consequence Argument as the key indirect challenge from casual determinism. I shall therefore not spend more time here relaying the standard structure of the argument again.

The Direct Challenges from causal determinism are worth briefly clarifying. An example of one of the Direct Challenges can be understood as the familiar Consequence Argument but restructured with a focus on moral responsibility:[64]

1. Nobody is morally responsible for the distant past or the laws of nature.
2. No one is morally responsible for the truth of determinism. In other words, no one is morally responsible for the fact that the distant past and the laws of nature entail our present actions and circumstances.
3. Therefore no one is morally responsible for their present actions or circumstances.

Let us call this example the Direct Argument from causal determinism. F&R note that the Direct Argument also makes the use of its own version of the transfer principle (1998, 24). We may refer to it as the 'non-responsibility transfer principle':

> *The non-responsibility transfer principle*: If we are not morally responsible for X or for X entailing Y, then we are not morally responsible for Y.

The Direct Argument poses a more direct threat to compatibilism between determinism and moral responsibility than the Indirect Challenges, as it does not argue for incompatibility between moral responsibility and determinism via the rejection of control or the

[63] For further examples and wordings of the Consequence Argument, see Fischer 1983, Ginet 1966 and Wiggins 1973.

[64] For another take on the Direct Challenge, see van Inwagen 1999.

ability to do otherwise in a determined system. Instead, thanks to its non-responsibility transfer principle, it argues that moral responsibility in itself is incompatible with determinism.

Recall that I argued in Chapter 2 that the machine incompatibilist argument is nothing more than the Consequence Argument applied to the specific case of machines. I shall take it for granted that any varieties of the machine incompatibilist argument can be fairly summarised in the same manner, as just machine-specific versions of either one of the Indirect or Direct Challenges to compatibilism.

F&R provide a carefully-laid-out account of moral responsibility using reasons-responsiveness, which includes a response to both the Indirect and the Direct Challenge. I will in this chapter present F&R's account alongside a running discussion of its suitability as the foundation for a machine compatibilist account.

First, F&R distinguish between two types of control in order to determine which type of control is necessary for the freedom-relevant condition for moral responsibility. F&R differentiate between guidance control and regulative control (1998, 31). Both of these types may be most easily introduced with an example. I will here use F&R's example of a woman, Sally, driving her car (F&R1998, 30 and Fischer 2012, 6). Sally's car has no mechanical problems and functions as expected. Imagine now that Sally has the intention to turn right in her car. As a result of her having this intention, she turns the steering wheel and guides the car to the right. In this setting, one might assume that Sally could have had the intention to turn left in her car instead. Had Sally formed the intention to go left, one might assume that she would have turned the steering

wheel and the car would have gone to the left. In this ordinary case, Sally controls the car and its movements as she guides the car to the right, but she could have guided it to the left instead.[65]

Using this example, F&R differentiates between two types of control. First, insofar that Sally guides her car to the right in the actual sequence of events, she has 'guidance control'. In the actual sequence of events that lead to Sally's car to going right, it is Sally who guides the car to the right. There is no intervention from an evil doctor, a pesky neural chip, or any other bothersome creation. Sally forms an intention and acts upon it. Sally turns the steering wheel and guides the car to the right. In short – in the actual sequence of events, Sally is in control of her car going to the right. In broad strokes, this is the kind of control that Fischer and Ravizza label 'guidance control'.

The second type of control is regulative control. Insofar that Sally also has the power to guide the car left, she has regulative control. As F&R put it: 'Regulative control involves a *dual* power' (1998, 31). Imagine the case featuring Sally and her car yet again. With the discussion of guidance control, the focus was set firmly on the actual sequence of events: the fact that Sally does what she intends to do, is not being interfered with by a pesky neural chip, and so on. Now, we shall shift the focus to a possible alternative sequence of events. Suppose that Sally had instead formed the intention to steer her car to the left. Assume that the possible alternative sequence of events that would have flowed from this involves no external interference, etc. In this alternative sequence, we can then assume that Sally acts on her intention and successfully guides her car to the left. In the

[65] I have here altered the wording of the Sally example to fit Fischer's newer recounts of the example, e.g. the one found in Fischer 2012. I have done this because the original wording included the notion of acting freely and was used to show that acting freely (as opposed to meeting the control requirement on moral responsibility) requires only guidance control – a view that F&R apparently subscribed to at the time of the example's invention. In Fischer's more recent work, this view has been omitted and his telling of the Sally example reflects this change in view. As the issues concerning full-blooded free action is not necessary for the purposes of this thesis, I shall here follow the authors' more recent approach and focus only on the control requirement on moral responsibility.

alternative sequence, then, Sally also has guidance control over her car. The 'dual power' that F&R refer to thus amounts to not only having guidance control over one's actual action (in Sally's case, turning right) but also having the power to do – and have guidance control over – some other, alternative action (turning left).

This control, involving both guidance control in the actual sequence of events as and in a possible alternative sequence of events, is what F&R call 'regulative control' (1998, 31). If the description of regulative control seems suspiciously familiar, there is good reason for this. Regulative control's extra demand for control in an alternative sequence may also be understood as the demand for an ability to do otherwise in a given situation. In other words, regulative control requires both guidance control in the actual sequence and the ability to do otherwise.

Given the example featuring Sally and her car, one might be excused for initially thinking that guidance control and regulative control mostly go hand in hand. However, Fischer and Ravizza use a Frankfurt-style case to aptly show that these two types of control do not necessarily appear together. Recall the Frankfurt-style case presented in Chapter 2. In the case in question, Jones intends and plans to kill another man, Smith. Recall further that a heinous character, Black, has ensured through some back-up plan or device that Jones is incapable of not deciding to kill Smith. As it happens, Jones decide and kills Smith without Black's fail safe ever coming into play (Frankfurt 1969, 835-836).

Jones forms the intention to kill Smith without any intervention from Black's side, and this intention leads him to his action of killing Smith. So, in the case described, Jones has guidance control vis-à-vis killing Smith, assuming his decision-making and acting are such as to fulfil F&R's more precise requirements for guidance control, which I will come onto shortly. However, while he does have control in the actual sequence of events, he could not have performed another action than

killing Smith: there is no alternative sequence of events that Jones is able to bring about, thanks to Black's present threat of interference. As such, Jones does not have regulative control regarding his action of deciding to kill and killing Smith.

As one might recall, Jones in the Frankfurt-style case is intuitively morally responsible for killing Smith. Fischer and Ravizza point out that this moral responsibility is attributed to Jones despite him not having regulative control with respect to his action of killing Smith. This conclusion is in line with the standard interpretation of Frankfurt-style cases, namely that they supposedly show that the ability to do otherwise is not necessary for attribution of moral responsibility.[66] F&R argue that Jones' moral responsibility for his action is explained by him having guidance control. They write: 'When we are morally responsible for our actions, we *do* possess a kind of control…. But the relevant sort of control need *not* involve alternative possibilities. The suggestion, derived from the Frankfurt-type cases, is that the sort of control necessarily associated with moral responsibility for action is *guidance control*' (1998, 32-33).

Hence, Fischer and Ravizza's account joins the chorus of compatibilist theories that rejects PAP.[67] Instead of focusing on alternate possibilities, Fischer and Ravizza recommend looking at the 'actual sequence' of events leading to a performance of an action in question to determine whether an agent exhibits guidance control and is therefore morally responsible regarding the action in question. They write the following reflection on Frankfurt-style cases: "…these cases invite us to develop what we shall call an 'actual-sequence' account of moral responsibility. By an 'actual-sequence' approach, we mean an approach to moral responsibility that does *not* require alternate possibilities… rather, what is important is (roughly speaking) what the agents actually do,

[66] Both PAP and Frankfurt-style cases were introduced in Chapter 2 of this thesis, but see also Robb 2020 for a detailed breakdown of their shared history.

[67] Also sometimes referred to as source compatibilists, see Levy and McKenna 2009.

and how their actions come to be performed" (Fischer and Ravizza 1998, 37). Thus – to return to Sally – if determinism is true, Sally can perfectly well be morally responsible for deciding to turn, and for actually turning, right, even though she lacks regulative control over that decision and action, because she still has guidance control, and that is enough to satisfy the freedom-relevant condition on moral responsibility.

F&R explain, then, how 'control' may refer to either regulative control or guidance control. The first of these, regulative control, is the type of control or 'freedom' that is generally referred to in the Indirect Challenges. As established, F&R reject the Principle of Alternate Possibilities (PAP). The ability to do otherwise, i.e. regulative control, is not a necessary component for moral responsibility on F&R's account. Instead, guidance control is the key to moral responsibility.

Hence, insofar as the Indirect Challenges claim to establish incompatibilism by appealing to the intermediary notion of regulative control, this type of challenge is not applicable to F&R's account. It may very well be that determinism is incompatible with regulative control, but if regulative control is not necessary for moral responsibility as contended by F&R, then the Indirect Challenges fail to establish the incompatibility of determinism and moral responsibility.

The next step, then, is to clarify what guidance control consists of in order to develop an actual-sequence account of moral responsibility, since so far we have simply taken it for granted that both Sally and Jones possess it with respect to their respective decisions and actions. F&R claim that two concepts are key when assessing whether an agent is exhibiting guidance control: "… an agent exhibits guidance control of an action insofar as the mechanism that actually issues in the action is his own reasons-responsive mechanism" (1998, 39).

The two concepts are therefore: reasons-responsiveness and ownership of the relevant mechanism. In the following sections, I will tackle these two concepts one after the other and demonstrate why F&R take them to be the essential core of guidance control.

Before I get that far, however, I shall briefly explain what F&R mean by 'mechanism'. In the quest to develop an actual-sequence account of moral responsibility, F&R shift the focus from agents to mechanisms. They write: "We contend that one very useful way to develop an actual-sequence approach to moral responsibility is to switch from a focus on the relevant *agents* and their properties, to a focus on the processes or 'mechanisms' that actually lead to the action" (1998, 38). As an example, for cases featuring human agents, this switch would usually mean a focal switch from the agent themselves and their abilities to the mechanism of their decision-making, such as their faculty for rational decision-making. I will discuss this focal switch later in the chapter in the section about mechanism ownership, but for now the above explanation will do.

II. Reasons-Responsiveness

As we just saw, Fischer and Ravizza take guidance control to have two components: reasons-responsiveness and ownership of the relevant mechanism. In this section, I explain reasons-responsiveness, which itself involves two central components: reasons-receptivity and reasons-reactivity (1998, 69-76).[68] I will then explain how, on F&R's account, reasons-responsiveness comes in degrees, by going over the two extreme points of the scale: weak reasons-responsiveness and strong reasons-responsiveness.

[68] Fischer and Ravizza are not the only compatibilists who link moral responsibility to a type of sensitivity to reasons, see Brink and Nelkin 2013, McKenna 2013 and Sartorio 2015 for a variety of other examples.

Consider the following story as an example of how these two components, reasons-receptivity and reasons-reactivity, come into play. Imagine a perfectly ordinary woman in a perfectly ordinary situation. The woman, Ruth, decides to buy and does buy an ice-cream cone from an ice-cream van at her local park. Ruth uses her faculty for rational decision-making in this actual sequence of events. Ruth has excellent reasons for buying herself an ice-cream cone: the weather is particularly delightful today, there are no queues at the ice-cream van and of course, Ruth has a desire to eat ice cream. Moreover, Ruth recognises these reasons – in other words, she is receptive to the given reasons. Based on these reasons, she decides to buy and enjoy an ice-cream cone. She is therefore not merely receptive to the reasons, but also reactive to them: they in fact motivate her to buy the ice cream (as they should). So, she is both receptive and reactive to reasons with respect to buying an ice-cream cone; hence she is reasons-responsive.

As I said above, F&R take reasons-responsiveness to apply not to the agent who performs the action, but to the mechanism that leads to the action. In the above scenario, the relevant mechanism would be Ruth's faculty for rational decision-making. The reasons-responsiveness of a mechanism does not depend merely on what happens in the actual sequence of events, which roughly in this case is that Ruth considers the pros and cons of buying an ice-cream cone and then decides to buy it. Instead, both reasons-receptivity and reasons-reactivity are determined by holding the relevant mechanism fixed and then considering what the mechanism does in various counterfactual scenarios. Precisely why we need to consider such counterfactual scenarios will become clearer later in this section, but for now the basic idea is simply that reasons-receptivity and reasons-reactivity are broadly dispositional concepts. This means that we may therefore not be able tell whether or to what extent they apply to the agent just by looking at what actually happens. For example, the agent might do the thing that there is sufficient reason to do – so

they are doing just what a reasons-responsive person would do – but be doing it for completely different (and bad) reasons, or for no reason at all.

An easy way of understanding these counterfactual scenarios is through the use of possible worlds.[69] Let us start by considering the operation of Ruth's relevant mechanism, her decision-making faculty, in the above scenario in just one other nearby world. Say that in this particular nearby world, the weather is rubbish. Furthermore, in this scenario the rubbish weather is sufficient reason for Ruth to not buy an ice-cream cone. Suppose that in this counterfactual world, where there is sufficient reason to not buy ice cream, Ruth recognises this reason to do otherwise and, on that basis, decides not to buy an ice-cream cone. Her recognising that reason is relevant to her reasons-receptivity and the fact that she factors that reason appropriately into her decision is relevant to her reasons-reactivity, as I explain below.

According to F&R, the fact that she is both receptive to, and reacts appropriately to, the relevant reason (the bad weather) in some possible world (namely the one just described) renders her actual mechanism 'weakly reasons-responsive': "under weak reasons-responsiveness, we… hold fixed the actual kind of mechanism, and then we simply require that there exists *some* possible scenario (or possible world) in which there is sufficient reason to do otherwise, the agent recognizes this reason, and the agent does otherwise" (1998, 44).

So, for Ruths's mechanism to be weakly reasons-responsive it must be the case that, holding fixed the actual features of that mechanism, there is at least one possible alternative scenario in which there is sufficient reason not to buy the ice cream (e.g. the weather is terrible)

[69] See Menzel 2021 for an introduction and overview of the history and usage of possible world semantics.

where she both recognises that reason not to buy it and then decides not to buy it.[70] Since Ruth's mechanism satisfies these conditions, it is weakly reasons-responsive.

The conditions for weak reasons-responsiveness are easy to satisfy, as there just needs to be *some* other possible world where the agent has sufficient reason to do, and does, otherwise. On the other end of the reasons-responsiveness scale lies strong reasons-responsiveness. To assess whether a mechanism is strongly reasons-responsive, one must look at not just one possible world, but instead all the nearby possible worlds where the mechanism operates and where there are sufficient reasons to do otherwise. The assessment is then made on whether the agent does otherwise in all these possible worlds.

F&R's definition of strong reasons-responsiveness, where K is the mechanism that actually issues an action, is as follows:

> "Strong reasons-responsiveness obtains under the following conditions: if *K* were to operate and there were sufficient reason to do otherwise, the agent would *recognize* the sufficient reason to do otherwise and thus *choose* to do otherwise and *do* otherwise" (1998, 41).

F&R deploy a counterfactual here: 'if K were to operate and there were sufficient reason to do otherwise ...'. In line with the standard Lewisian semantics for counterfactuals, we can analyse 'if A were to be the case, then B would be the case' as 'in all the closest possible worlds where A is the case, B is the case' (Lewis 1973). Thus understood, F&R are saying that strong reasons-responsiveness requires the following: in all the closest possible worlds where K operates and there

[70] Holding the mechanism fixed is just a technical addendum used to avoid problems with Frankfurt-style cases. I shall return to this in a later section.

is sufficient reason to do otherwise, the agent recognises that reason and thus chooses to do, and does, otherwise.

Now, consider what Ruth would have done had the weather been rubbish, the queue for the ice-cream van gone around the block, or if she was on a tight budget that did not allow for frivolous purchases of ice cream. In all these scenarios, there is sufficient reason for Ruth not to buy an ice-cream cone. Suppose that these are all the closest possible worlds, where there was sufficient reason to not buy ice cream. And suppose, further, that in all of those worlds Ruth, using her faculty for rational decision-making, recognises these reasons and decide to not buy an ice-cream cone. Then, Ruth's faculty for rational decision-making is strongly reasons-responsive vis-à-vis buying an ice-cream cone, because had she had sufficient reason to do otherwise, she would have recognised that reason and duly done otherwise.

For a mechanism to be strongly reasons-responsive with respect to an action the actual world, three key things must happen in all the closest possible worlds, where the mechanism operates and there is sufficient reason to do otherwise. There must be a recognition of sufficient reason to do otherwise, a choice made based on the reason, and the act performed accordingly. F&R hence identify three corresponding types of failures in alternatives sequences of events, which would keep an agent from being considered strongly reasons-responsive (1998, 41). I will briefly go over these three failures here in order to make it clear why reasons-responsiveness is an amalgamated concept made up of two independent conditions: reasons-receptivity and reasons-reactivity.

The first failure is the failure to be strongly receptive to reasons.[71] This failure is commonly found in scenarios featuring agents acting under delusions, hypnosis or similar. For example, imagine a woman, Anna, whose situation in the actual world is reads completely identical to Ruth's situation from earlier. Anna also recognises the good weather as a reason to buy an ice-cream cone and so decides to buy one. She similarly uses her faculty for rational decision-making. As far as the actual sequence of events is concerned, Anna seems completely identical to Ruth.

Here is how the two women differ, however. As we wish to assess whether Anna is strongly reasons-responsive, we'll look at how she fares in a nearby possible world where her mechanism operates and there is sufficient reason to do otherwise. Suppose in this possible world, the queue for the ice-cream van is miles long. This is a sufficient reason for Anna to not buy an ice-cream cone, as she does not have the time to spare to stand in a queue that long. Then, in this other possible world, Anna has a sufficient reason to not buy an ice-cream cone, i.e. to do otherwise. Yet, suppose that in this world Anna, like Ruth, does choose not to buy an ice-cream cone, but, unlike Ruth, not for this given reason. Instead, Anna suffers from paranoid delusions. As part of these paranoid delusions, Anna is made to believe that the people queueing are all secretly spies, who are just waiting to get rid of her. She fears for her life, and genuinely thinks that joining the queue would mean her death. For this reason, and this reason only, she chooses to not buy an ice-cream cone.

Notice that Ruth and Anna's actual situations are the same: the same reasons and the same decision-making. Anna is not delusional in the actual scenario, as her delusion is only triggered by the long queue in the nearby possible world. In that possible world, Anna fails to

[71] See also Duggan and Gert 1979 for a discussion of this type of alternative-sequence failure. See Kozuch and McKenna 2016 for a more nuanced discussion of the relationship between failure to be reasons-responsive and mentail illness.

recognise the long queue time as a sufficient reason to do otherwise. Instead, she only sees her imminent imagined death as a reason to not buy ice cream. Anna's decision-making mechanism therefore fails to be strongly receptive - meaning that it is not true that, were there to have been sufficient do otherwise, she would have recognised those reasons. The above is then an example of how a mechanism can fail to be reasons-receptive. This constitutes the first of the three possible ways in which a mechanism can fail to be strongly reasons-responsive.

While the first type of failure is about reasons-recognition, the second and third types of failure are both failures to be reactive to reasons. The second type of failure occurs when the agent in a relevant nearby possible world recognises that there is sufficient reason to do otherwise, yet he chooses not to act accordingly. So, the failure is a failure to move from recognising reasons to do otherwise to choosing to actually act in accordance with these reasons. The third type of failure occurs when the agent in a relevant nearby possible world recognises that there is sufficient reason to do otherwise and chooses to do otherwise, yet fails to act upon this choice. The failure here is therefore a failure to move from making a choice based on sufficient reasons to do otherwise, and actually acting in accordance with this choice. These are the last two types of failure which can keep mechanisms from being considered strongly reasons-responsive. While these last two types of failures are distinct, I will for the purposes of this thesis refer to them both as failures regarding reactivity to reasons. One may easily create an example of a mechanism failing to be reasons-reactive due to weakness of the will.[72]

Imagine a woman, Beatrice, whose actual sequence of events is exactly the same as that of Ruth, our strongly reasons-responsive agent. Beatrice also decides to buy and does buy an

[72] Weakness of the will is a commonly used example within philosophical contexts and is generally understood as actions performed by an agent against her better judgement, see for examples Davidson 1980 and Mele 1995.

ice-cream cone from an ice-cream van at her local park, uses her faculty for rational decision-making in the actual scenario, and so on. Now, just like in Anna's case, we will investigate whether Beatrice is strongly reasons-responsive with respect to her ice cream buying by looking at nearby possible worlds where her mechanism is still operating, and she has reason to do otherwise. This is where Beatrice differs from both Ruth and Anna. Suppose that in a nearby possible world, Beatrice's faculty for rational decision-making operates and she has sufficient reason to do otherwise. Suppose the reason not to buy an ice-cream cone in this world is that Beatrice is on a student budget, and there is no space in the budget for buying ice-cream cones. Beatrice recognises this as a sufficient reason to not buy an ice-cream cone, i.e. to do otherwise. Nevertheless, she suffers from weakness of the will at that moment. She chooses to buy an ice cream, despite her budget concerns.

In this example then, Beatrice does recognise the sufficient reason to do otherwise, yet fails to translate this recognition into reacting accordingly. This is therefore an example of a failure regarding reasons-reactivity, and her mechanism therefore fails to be strongly reasons-responsive. [73]

It should now be clear what defines strong reasons-responsiveness, weak reasons-responsiveness, and how an agent, such as Anna and Beatrice, can fail to be strongly reasons-responsive despite acting in accordance with sufficient reasons in their actual sequence of events. I shall now briefly explain why, according to F&R, neither extreme of reasons-responsiveness is useable as the control (or freedom-relevant) condition on moral responsibility.

[73] Other example cases featuring reactivity failures includes phenomena such as phobias, compulsive neurotics or physical incapacites, see Fischer and Ravizza 1998.

To start, strong reasons-responsiveness is too limiting as a condition, as it excludes many kinds of case where attribution of moral responsibility is intuitively appropriate.[74] This includes agents whose relevant mechanism in just one nearby possible world suffers from weakness of the will, mental challenges (including phenomena such as phobias) and similar. I will here demonstrate with some simple examples.

Imagine that an agent, May, decides to spend her evening volunteering for a charity, and based on this decision she follows through and does the volunteering. Suppose that May makes this choice by using her well-functioning faculty of rational decision-making. In this case, there is no intervention or impairment to her decision-making. To assess whether May is strongly reasons-responsive, we must consider the closest possible worlds where May has sufficient reason to do otherwise than volunteer.

Say that in a nearby possible world, May has an imminent deadline for some university coursework. She has, in this alternative possible world, sufficient reason not to do her volunteering that day, but instead to do her coursework. We may suppose that she recognises this to be a sufficient reason not to volunteer. Yet suppose that, despite this, May still decides to show up to her volunteering. In this possible world, then, May fails to be appropriately affected by her belief that there is sufficient reason for her to stay at home and do coursework. Suppose, further, that in all other relevant nearby possible worlds, where May has some sufficient reason not to volunteer – say, worlds where a close friend urgently needs her help, or she tests positive for coronavirus – she does not show up to her volunteering. Nonetheless, May cannot be considered strongly reasons-responsive with respect to her decision to do her volunteer work in the actual sequence of events,

[74] Though some argue that using anything less than strong reasons-responsiveness for guidance control on Fischer and Ravizza's account renders the account too lenient, see for example Stout 2016.

because of this one failure to react to a sufficient reason to do otherwise in just a single relevant possible world. This should be a clear example of how constricting the criteria for strong reasons-responsiveness are. One failure in just one nearby relevant possible world is enough to disqualify a mechanism from strong reasons-responsiveness.

This consequence runs counter to ordinary moral practice. After all, as members of a moral community, we would surely want to commend May for her volunteering, thus seeing her as responsible for her own good deed. Since an adequate account of moral responsibility ought to be able to capture and explain our uncontroversial moral practices, strong reasons-responsiveness is too limiting to be a necessary condition for moral responsibility.

On the other end of the spectrum, F&R argue that weak reasons-responsiveness is too lenient to be a sufficient or co-sufficient for moral responsibility, because it can be attributed to agents in cases in which they clearly should not be held morally responsible.[75] That the condition is too lenient becomes especially obvious when considering cases featuring agents whose behaviour manifests in strange patterns. As an example, consider the following case. First, recall May who volunteers for her local charity. Imagine now May to instead be of a most curious nature.

In the actual sequence of events, May is gripped by a sudden, overwhelming an, even to her, inexplicable urge to set the charity headquarters on fire. While there are obvious sufficient reasons not to do so, the nature of the urge is such that May is simply oblivious to those reasons: she simply does not recognise the damage, suffering, risk of serious harm and waste the emergency

[75] An earlier version of F&R's 1998 account, promoted by Fischer 1987 and 1994 focused on weak reasons-responsiveness. Multiple objections were raised against this account on the back of the leniency of weak reasons-responsiveness; particularly noteworthy were van Inwagen 1997 and Schoeman's Saber-slayer thought experiment mentioned in Fischer and Ravizza 1998, p.65.

services' scarce resources as reasons no to light the fire. Indeed, in nearly all possible scenarios there is a sufficient (indeed, overwhelming) reason not to commit arson, yet she sets the headquarters on fire – all, that is, bar one.

If May were to see a moth pass the patio light outside, it would trigger an old memory that would shake May to her senses, see the sufficient reasons to refrain from arson for what they are, and instead she would just do her normal volunteer work. The existence of this one nearby possible world where the relevant mechanism operates and May acts on a sufficient reason not to burn down the headquarters implies that May's decision-making mechanism is weakly reasons-responsive vis-à-vis her act of setting the charity building on fire.

May is weakly reasons-responsive, but it seems incredibly implausible that we would consider her morally responsible for setting the building on fire. She was, after all, acting on the basis of a sudden, irrational and inexplicable urge. Furthermore, the one possible reason for not committing arson which she could recognise in any relevant possible world involved the flight pattern of a moth. If anything, the casual moral judge would deem her to lack the sanity to be considered morally responsible.

It should now be evident how there is an obvious problem with using weak reasons-responsiveness as a sufficient condition for moral responsibility. It would count agents such as May, whose mechanisms display strong erratic behaviour and patterns of nonsensical reasoning, as morally responsible. So where strong reasons-responsiveness is too restrictive as a condition for moral responsibility, weak reasons-responsiveness is too lax. For these reasons, F&R argue that neither extreme of reasons-responsiveness is fitting as a condition for moral responsibility. Instead, a middle ground must be found.

III. Moderate Reasons-Responsiveness

In this section, I will explain how F&R use a 'middle way' between the two extremes of reasons-responsiveness, namely moderate reasons-responsiveness, to develop an account of the control condition for moral responsibility, which captures our intuitions about ordinary moral practices.

As discussed in the previous section, on one hand the combination of strong receptivity and strong reactivity, i.e. strong reasons-responsiveness, is too demanding to deliver a satisfactory account of guidance control, and hence to play a part in the conditions for moral responsibility. On the other hand, the combination of weak receptivity and weak reactivity, i.e. weak reasons-responsiveness, is much too lenient to be used for guidance control. The required middle ground, F&R argue, is to be found in an asymmetry between receptivity and reactivity: "We contend that the reactivity to reasons and receptivity to reason that constitutes the responsiveness relevant to moral responsibility are crucially *asymmetric*. Whereas a very weak sort of reactivity is all that is required, a *stronger* sort of receptivity to reasons is necessary for this kind of responsiveness" (1998, 69).

A stronger sort of receptivity and a weak sort of reactivity hence constitute what F&R call moderate reasons-responsiveness. I will in this section explain what exactly is being referred to when speaking of stronger and weaker sorts of receptivity and reactivity, as well as show how F&R defend this asymmetry claim. I will start by clarifying F&R's position on the reactivity to reasons needed for guidance control, and then move on to discussing their take on the necessary receptivity to reasons. For this purpose, using F&R's own example is the most illuminating.

They ask the reader to imagine a weak-willed individual named Brown (F&R 1998, 69). Brown is an avid user of a non-addictive drug called Plezu. Plezu does not cause irresistible urges to take it and has no withdrawal symptoms for its users. As a drug, it instead just activates the

brain of its user to cause an intense experience of euphoria. The effects of the drug lasts for most of a day, and under the influence the user is unable to do anything but lie down and revel in the euphoria. As such, unrestrained or even merely regular use of the drug usually costs the users their job, family and friends.

Brown, our weak-willed individual, recognises that there are more than sufficient reasons for him to not take Plezu so avidly. Despite this, he spends most of his days under its influence: lying down, lost in euphoria. Suppose that there exists only one nearby possible world, where Brown chooses not to take Plezu. Imagine in this particular possible world, Plezu has been found to kill its avid users – a most horrendous and unfortunate side effect. In this nearby possible world, Brown is informed that if he takes Plezu one more time, it will be his death. Say then that in this one nearby possible world, Brown reacts to this sufficient reason for not taking the drug and as such does otherwise. As Brown only does otherwise in this one nearby possible world, the relevant mechanism (that is his faculty for rational decision-making) only qualifies as being weakly reasons-reactive.

F&R argue that the fact that Brown does not take Plezu in at least one nearby possible world reveals two things. First, it proves that Brown's urge to take Plezu is not irresistible, as he does refrain from taking it in one nearby possible world by reacting to relevant reasons. Second, Brown's refraining from taking Plezu in one possible sequence of events delivers sufficient reasons-reactivity for guidance control, and in turn moral responsibility.

This second claim deserves further explanation. F&R's statement is based on the claim that 'reactivity is all of a piece': "*holding fixed the actual kind of mechanism*, reactivity is all of a piece: if the mechanism can react to any reason to do otherwise, it can react to all such reasons" (F&R 1998, 74). There are two core concepts behind this claim. Take a mechanism operating in a

given scenario. First, if that very mechanism reacts (and hence *can* react) to a reason to do otherwise in another possible world, then it can also react to a reason to do otherwise in the actual world. Since the mechanism is held fixed between the worlds, if it has a given capacity in one world, it has same capacity in all worlds. Second, the capacity in question should not be thought of as merely the capacity to react to a *specific* sufficient reason to do otherwise (in this case, the fact that taking the drug will result in death). Rather, it is the *general* capacity to react to *any* sufficient reason to do otherwise. So even though Brown in fact fails to react to a sufficient reason to do otherwise (namely loss of jobs, friends, and so on), he still has the *capacity* to react to that sufficient reason, since he has the general capacity to react to any sufficient reason, including the one that actually obtains.[76] In this way, F&R argue that weak reactivity is sufficient for the reactivity component of the reasons-responsiveness required for moral responsibility – that is, for moderate reasons-responsiveness.

Let us now turn our attention to the level of receptivity to reasons needed for moderate reasons-responsiveness. As mentioned, F&R put forward an asymmetry claim regarding the needed reactivity and receptivity to reason needed to ground guidance control. While only weak reactivity to reasons is required, weak receptivity to reasons is not enough. To illustrate why, recall the example of May and her arsonist ways from the previous section. In the example, May is weakly reasons-receptive, as it is only in the one possible world, where a moth flying past triggers a specific memory for her, that she recognise a sufficient reason not to burn down the charity headquarters. But it is consistent with this that there is nothing wrong with May's capacity to *react* to reasons, once she recognises them; after all, in the single possible world where she recognises

[76] This is a fairly controversial aspect of F&R's 1998 account. For discussions, see Davenport 2002, Ramirez 2015 and McKenna 2016. I will in the following chapter show why objections to this this aspect of F&R's account do not influence the machine compatibilist version of the account, when applied to autonomous systems.

them, she also reacts appropriately to them. Yet, May is not morally responsible for burning down the building. Therefore, as far as reasons-receptivity is concerned, guidance control must require more than *weak* reasons-receptivity.

F&R instead propose that the receptivity needed for moderate reasons-responsiveness is 'regular receptivity', which requires an understandable pattern of receptivity to reasons:

> "Regular reasons-receptivity, then, is reasons-receptivity that gives rise to a minimally comprehensible pattern, judged from some perspective that takes into account subjective features of the agent (i.e. the agent's preferences, values, and beliefs) but is also *not simply* the agent's point of view…. Regular receptivity to reasons, then, requires an understandable pattern of reasons-recognition, minimally grounded in reality" (F&R 1998, 73).

A moderate reasons-responsive mechanism must then be weakly reasons-reactive, as well as regular reasons-receptive – meaning it must exhibit a sensible pattern when it comes to recognising reasons. We can now create an example featuring a moderately reasons-responsive agent, which will showcase how this asymmetry between reasons-reactivity and -receptivity plays out in practice. Recall the ice-cream cone example from earlier, which featured our strongly reasons-responsive agent, Ruth. Imagine a woman, Charis, whose actual sequence of events runs completely identical to Ruths' situation. Charis, just like Ruth, recognises the sunny weather as a reason to buy an ice-cream cone, decides and does buy one. Just like Ruth, Charis uses her faculty for rational decision-making in this scenario.

To assess Charis' reasons-responsiveness with respect to her purchase, let us then consider some relevant possible worlds. Suppose that there is a range of nearby possible worlds, in which Charis' mechanism operates, and where there is a queue for the ice-cream van of different

lengths: 10 metres, 20 metres, 30 metres and so on. Imagine now, that in just one of these nearby possible worlds – say the one where the queue is 30 metres long – Charis both recognises the queue length as a reason to do otherwise and also acts upon this reason, thereby doing otherwise from the actual sequence of events. Since she does otherwise in just one of these relevant possible worlds, her mechanism is at least weakly reasons-reactive.

Though she might only do otherwise in this one possible world, suppose that Charis nevertheless recognises the queue length as a reason to not buy an ice-cream cone in all the mentioned relevant possible worlds. As such, though she is only weakly reasons-reactive – she only reacts appropriately to the queue as a reason to do otherwise in the 30-metre world, and not in the 10- or 20-metre worlds – she is still receptive to the long queue being a reason to do otherwise in the mentioned possible worlds where the queue is 10, 20, 30 metres and so on. Charis' recognition of the various queue lengths as reasons to do otherwise constitute a sensible pattern of reasons-recognition. After all, one may fairly assume that if an agent recognises a 10-metre-long queue as a sufficient reason not to get ice cream, then the same agent, if sensible, would logically also recognise the queue being longer than 10 metres as sufficient reason not to buy ice cream. Considering that Charis' mechanism exhibits precisely this sensible pattern, the mechanism is regularly reasons-receptive. Together with the mechanism being weakly reasons-reactive, Charis' mechanism fulfils the conditions for moderate reasons-responsiveness with respect to her ice-cream cone purchase.

Now, one might be tempted to think that the above example just shows Charis to be strongly reasons-receptive like Ruth, but this would be a mistake. Suppose that there are other relevant possible worlds, where Charis' mechanism operates, and where there are sufficient reasons to do otherwise that do not relate to the queue length. Suppose, for example, that in one world the

weather is horrible or that Charis is on a tight budget. Even if Charis is not receptive to these reasons (or any other imaginable ones), she is still regularly reasons-receptive thanks to her sensible display of reasons-receptivity with respect to queue length. Hence, the criteria for regular reasons-receptivity are far more relaxed than those for strong reasons-receptivity, where the mechanism must be receptive to every reason to do otherwise in all nearby possible worlds, where the mechanism operates.

F&R's account of moderate reasons-responsiveness, then, requires only weak reasons-reactivity, but regular reasons-receptivity. Moderate reasons-responsiveness is then more demanding than weak reasons-responsiveness, while still being fairly attainable compared to strong reasons-responsiveness. F&R summarise the receptivity and reactivity conditions for moderate reasons-responsiveness as follows:

"In the case of receptivity to reasons, the agent (holding fixed the relevant mechanism) must exhibit an understandable pattern of reasons-recognition, in order to render it plausible that his mechanism has the 'cognitive' power to recognise the actual incentive to do otherwise. In the case of reactivity to reasons, the agent (when acting from the relevant mechanism) must simply display *some* reactivity, in order to render it plausible that his mechanism has the 'executive power' to react to the actual incentive to do otherwise" (Fischer and Ravizza 1998, 75).

Now, recall the format of the Direct Challenges as it was laid out earlier in this chapter. In these, the incompatibilist conclusion does not rely on the intermediary notion of control or similar ideas, but instead argues for direct incompatibilism between determinism and moral responsibility. We can now begin to see how F&R's account seeks to answer the Direct Challenges.

Guidance control, being the key to moral responsibility on F&R's account, was shown to consist of two key components: moderate reasons-responsiveness and mechanism ownership.

F&R's answer to the threat posed by the Direct Challenges comprises of showing that each of these concepts are compatible with causal determinism.

For now, I shall talk about moderate reasons-responsiveness and determinism, while leaving mechanism ownership for the next section. For an agent to be moderately reasons-responsive, what is required is that various nearby possible worlds, in which (a) the agent's action is produced by the same mechanism as it is in the actual world and (b) there are sufficient reasons for the agent to do otherwise, have various features: the agent must, throughout those worlds, display an 'understandable pattern of reasons-recognition' ('regular reasons-receptivity'), and there must be *some* such worlds where the agent does otherwise ('weak reasons-reactivity').

Crucially, moderate reasons-responsiveness is compatible with determinism. This is simply because nearby possible worlds that meet conditions (a) and (b) above need not – and if determinism is true, will not – have exactly the same past and laws of nature as the actual world. While it may be true that Charis, our ice-cream-cone-buying agent, cannot do otherwise than buy the ice-cream cone, determined as she is to do so by the past and the laws, it is nonetheless the case that in at least some worlds where she has sufficient reason to do otherwise (namely worlds where the queue for the ice-cream van is very long), she does otherwise. Thus she counts as weakly reasons-reactive. Similarly for regular receptivity: even if an agent is determined by the past and the laws to act as they do in the actual world, that is entirely consistent with their acting differently in various other possible worlds where things are a bit different, in such a way that the agent displays, across those worlds, the required pattern of reasons-recognition. Hence, moderate reasons-responsiveness is compatible with determinism and can make up part of F&R's answer to the incompatibilist Direct Challenges.

As I said earlier, guidance control – the key to moral responsibility on F&R's account – consists of two key components: moderate reasons-responsiveness and mechanism ownership. I shall now turn my attention to the second element of guidance control: mechanism ownership.

IV. Mechanism Ownership

Moderate reasons-responsiveness, F&R claim, is not sufficient for guidance control by itself, because an agent could be moderately reasons-responsive and yet still be manipulated in such a way as to exempt them from moral responsibility. F&R therefore add a second component to guidance control: mechanism ownership. In short, F&R argue that for an agent to have guidance control vis-à-vis a given action, the relevant operating mechanism must be the agent's own. In this section I will attempt to explain what mechanism ownership amounts to, and why F&R argue that it is a necessary part of guidance control. I will further argue that this aspect of F&R's account is currently unable to answer some of the more recent manipulation worries.

To start, I will explain a bit about F&R's switch to a mechanism-focused rather than an agent-focused approach and its implications, as promised in §II. To start, recall the Frankfurt-style case featuring Black and Jones, as it was introduced back in Chapter 2. In this case, the agent of primary interest is Jones, while the mechanism in play in the actual sequence can be identified as Jones' faculty of practical reasoning, since Jones' action in the actual sequence stems from his practical reasoning without interference from Black.

In contrast, consider what would have happened had Jones been about to decide to not kill Smith. In that alternative sequence, Black's failsafe is triggered and ensures that Jones kills Smith. In terms of agents, little has changed. Jones is still the agent who performs the action of

killing Smith. However, the relevant mechanism in this alternative sequence has changed. The action stems not from Jones' practical reasoning, but rather from Black's failsafe mechanism; hence, it is this fail-safe mechanism that produces Jones' decision. This is an example of how a mechanism-based approach can render a different result from an agent-based one when dealing with Frankfurt-style scenarios.

When assessing the reasons-responsiveness of a mechanism, F&R furthermore requires that the relevant mechanism is held fixed, when considering various possible worlds, as described in the previous section. I will here show why this requirement is crucial, when dealing with a particular problem regarding Frankfurt-style cases on F&R's account. Specifically, the problem is that intuitively Jones is morally responsible for killing Smith, and yet he seems to fail to be moderately reasons-responsive in respect to killing Smith, as one might think that he fails to be weakly reasons-reactive: in *all* of the nearby possible worlds (including the actual world) in which there is sufficient reason not to kill Smith, he nonetheless kills Smith.

F&R's definition of reasons-responsiveness in terms of mechanisms rather than agents solves this worry. It is *not* true that in all nearby possible worlds *where Jones's action-producing mechanism is held fixed*, he fails to do what there is sufficient reason to do, because those possible worlds exclude the worlds in which Jones's action is produced by Black's failsafe device. There are, plausibly, reasonably nearby worlds – worlds where Black is absent or his failsafe device malfunctions – where Jones's practical reason delivers what there is sufficient reason to do, namely refrain from killing Smith. After all, if there weren't any such worlds, Black would hardly have needed to go to the bother in setting up his failsafe. Indeed, that there was a serious risk that Jones would, if left to his own devices, end up doing what there was sufficient reason to do – that is, decide not to kill Smith – was the very reason Black went to all that trouble in the first place.

The mechanism focus in F&R's account thus allows it to avoid running into trouble in cases featuring more complex impaired agents, such as those found in Frankfurt-style cases.[77] For the machine compatibilist project, this switch will become useful as it allows us to switch from considering autonomous systems as potential agents to focusing instead on their decision-making mechanisms.

Now, when it comes to F&R's concept of mechanism ownership, this is tied to their contention that moral responsibility is a historical notion. They write: 'The history of (say) an action is important *in part* because it helps to specify what it is for a mechanism to be the *agent's own*' (F&R 1998, 170). For the purposes of this section, I will make do with a simple and intuitive distinction between historical and non-historical notions. I will summarise the distinction here as follows (1998, 171):

> Non-historical phenomenon: a non-historical concept or thing is dependent only on properties that do not hinge on facts about the given concept or thing's history.[78]
>
> Historical phenomenon: a historical phenomenon is dependent on properties that are responsive to changes in the given concept or thing's history.

This distinction is fairly straight forward. Imagine first a small tin can, metallic and cylindrical. These particular properties, of being metallic and cylindrical, are non-historical features of the can. We do not need to know the tin's history to find out that it has these properties, nor would changes

[77] There is quite a bit more to be said on the topic of mechanisms versus agents, however that discussion will not be conducive to our purposes here. See Ginet 2006, Wallace 1997, McKenna 2001, Watson 2001 for further discussions of the use of the mechanism approach in F&R's account. See also Fischer's replies to these in, respectively, Fischer 2006, 2012, and (for the last two) 2004.

[78] Also referred to by F&R as 'snapshot' properties.

to the tin's backstory change the fact that the tin is metallic and cylindrical. As such, these properties are disconnected from the respective object's history.

By contrast, suppose then that this tin can is a misplaced piece of modern art - more specifically, it is part of Piero Manzoni's infamous work 'Merda d'artista'. These two features of the tin can – being a piece of modern art, and being created by Manzoni – are features that are only revealed in reference to the object's history. As such, these properties are historical features of the tin can, as we require knowledge of the can's history in order to ascribe these features, and a change to the tin can's history could change the fact that it has these properties.

F&R argue that moral responsibility is a historical phenomenon (1998, 195).[79] Specifically: "the past must contain a process of 'taking responsibility.' Taking responsibility, we believe, is a necessary feature of moral responsibility. It is part of the process by which a mechanism leading (say) to an action, becomes *one's own*." (F&R 1998, 207).

F&R contend that a vital part of any human's moral education is the slow and learned realisation and understanding of oneself as an agent. Through a standard moral education, a child learns about how her exercise of agency affects other beings, and is further encouraged to see herself as an appropriate target for reactive attitudes. As F&R puts it: "This sort of training aims to induce a certain sort of view in the child, a view of himself as an agent and, in some situations, a fair target for praise and blame" (1998, 210). Through moral 'training', a child learns to take responsibility for their doings and eventually is ready to be held responsible by the people around them as well.

[79] For more on moral responsibility as a historical versus a non-historical notion, see also Christman 1991, Vargas 2006 and D. Zimmerman 2003.

F&R argue that there are three conditions that must be fulfilled on their account for an agent to take responsibility. They are:

I. "an individual must see himself as the source of his behaviour in the sense we have specified. That is, the individual must see himself as an agent… The agent thus sees that his motivational states are the causal source – in certain characteristic ways – of upshots in the world." (F&R 1998, 210-211).

II. "the individual must accept that he is a fair target of the reactive attitudes as a result of how he exercises this agency in certain contexts." (F&R 1998, 211).

III. The individual's view of himself specified in the above conditions must be grounded in his evidence for these beliefs (F&R 1998, 213).

Now the pieces fall into place. Going back to their talk of moral education, F&R write: "By the time the child becomes a full member of the moral community, he is expected to have fully taken responsibility for the actions that flow from his mechanism of practical reason. In this sense, the mechanism of practical reason is now appropriately considered his *own*" (1998, 215)

Through taking responsibility, an agent therefore recognises a mechanism, such as their mechanism for practical reasoning - or even non-reflective mechanisms - as their own. Now it may begin to become clearer why F&R put such weight and importance on the notion of 'taking responsibility'. In reference to an agent's history, including his moral education, it is clear that an agent also does not and will not take responsibility for all of his possible actions, regardless how they come to be. F&R summarise it as such: "When an agent takes responsibility… he is accepting responsibility for only those actions which flows from a certain source. This idea can be framed

more precisely by saying that 'an agent takes responsibility for acting from a particular kind of mechanism" (1998, 215).[80] The particular kind of mechanism is, of course, the agent's own.

So far, the only examples of mechanisms have been agents' faculty for rational decision-making and Black's strange action-inducing device, and as such one might naturally question what other kinds of mechanisms exists and how one might individuate them from each other. F&R write: "we rely on the intuitive judgement that the normal mechanism of practical reasoning is different from deliberations that are induced by a significant direct electronic manipulation of the brain, hypnosis, subliminal advertising, and so forth' (1998, 40). As such, one might imagine Black's device, hypnosis and similar such things as mechanism capable of producing action, however these kinds of mechanisms are types that an agent does not usually take responsibility for.

Through an agent's history and self-view, he comes to take responsibility for certain mechanisms, such as his own faculty for rational decision-making, and recognise these as his own. As such, were an agent to be manipulated through hypnosis or similar things, he would not have a history of taking responsibility for the manipulated mechanism, and if informed about the manipulation, the agent would not recognise the manipulated mechanism as his own.

The mechanism aspect of F&R's account should now stand clear. It was shown much earlier in this chapter that using mechanisms as a focal point of the account allowed F&R to bypass worries stemming from Frankfurt-style cases. Here, it was explained how mechanism ownership can be established through the given agent's history of taking responsibility. So, the establishing of moral responsibility as a historical notion, of the importance of taking responsibility being part of

[80] F&R are not alone in pursuing an account of taking responsibility: see Enoch 2012, Mason 2019 and Wolf 2001 for different takes on the concept of taking responsibility.

moral education and so on, all boils down to a simple thing: how a mechanism can be appropriately said to be an agent's own, in order to avoid worries about such things as manipulation. This then makes up the mechanism ownership part of F&R's account.

Now, we may also see how F&R's account seeks to respond to the Direct Challenges, which one might recall from the first section of this chapter. In the previous section, I mentioned that F&R's answer consists of showing that both components of guidance control - moderate reasons-responsiveness and mechanism ownership – are compatible with determinism. I explained at the end of the previous section how moderate reasons-responsiveness is compatible with determinism. Mechanism ownership is also clearly compatible with determinism. An agent may be determined and still fulfil the three conditions for taking responsibility mentioned earlier. The process of moral development by which agents come to take responsibility for, and thus own, an action-producing mechanism is not at odds with determinism as a thesis. As such, F&R answer the Direct Challenge, by claiming that guidance control is compatible with determinism.

Unfortunately, however, given the attention that manipulation cases have attracted in recent years, it is not completely clear that F&R's account successfully avoids both the Direct and Indirect challenges.[81] The original intuitive worry about determinism formulated in the Indirect Challenges was that the seemingly lack of freedom or ability to do otherwise might be incompatible with moral responsibility. Manipulation cases pose a similar question. Intuitively, a manipulated agent is usually not morally responsible for their manipulated actions. After all, our ordinary moral practices generally do not take agents who are involuntarily hypnotised, forced at gunpoint or otherwise manipulated to be morally responsible. However, for many incompatibilists, there is a

[81] See McKenna and Coates 2021, Mele 2008 and McKenna 2012 for more on how manipulation arguments present a serious challenge for contemporary compatibilism.

striking similarity between the ordinary determined agent and the manipulated agent: at surface level at least, they both seem to fail to meet the freedom-relevant condition for moral responsibility.[82]

As compatibilists leave PAP behind and propose conditions for moral responsibility that do not require the ability to do otherwise, a question arises. Can proponents of these new compatibilist theories, who claim to have successfully shown how determined agents can nonetheless be morally responsible agents, also concurrently explain why manipulated agents are not morally responsible? The mechanism ownership aspect of F&R's reasons-responsive compatibilism is supposed to do this job, but it is arguably the most underdeveloped part of their account. Hence, the account struggles to deal with a range of relevant manipulation worries.

One such worry is succinctly expressed by Levy (2011), utilising the famous Ann/Beth thought experiment by Mele (1995, 145-146). Imagine two philosophy professors, Ann and Beth. Both are doing well at their job, but Ann is much more productive and publishes a great deal more than Beth. Their dean, firmly caught in a publish or perish mentality, hires neuropsychologists to rewire Beth as she sleeps. Following the dean's wishes, Beth is rewired such that her values towards philosophy, hard work, leisure and family are all changed to match Ann's. This rewiring, though seemingly drastic, is not in any danger of threatening the continuity of Beth's personal identity.[83] Next morning, Beth finds herself with a changed outlook on her work/life balance and wanting to dedicate herself completely to the study of philosophy. While Beth is surprised by this, she wholeheartedly endorses it, as she finds it in complete alignment with her new

[82] See Taylor 1974, Kane 1996 and Watson 1999 for different variations of this worry and for examples of manipulation cases.

[83] For a discussion of manipulation and personal identity in relation to compatibilism, see Matheson 2014.

altered values. Contrary to her usual schedule, Beth starts the morning by immediately working on a new paper.

Both Mele (1995) and Levy (2001) claim that intuitively Beth cannot be praised nor held morally responsible for her action. However, in contrast, Levy (2001, 104-105) argues that on F&R's account, Beth can take responsibility for her actions and be morally responsible, despite having undergone some serious manipulation at the hand of neuropsychologists. Levy suggests that Beth could fulfil the mechanism ownership condition with respect to her action by merely reflecting upon her new values before getting up to write her paper. Beth does not know about the neuropsychologists doing work on her, but she does recognise that her altered values are new. On previous mornings, she recognised herself as a normal agent who is normally the source of her own actions, and saw herself as a fair target of reactive attitudes. Though her altered values are new, on this morning she would see no reason to think that any of this should change. So, since Beth satisfies the ownership condition (and she is also moderately reasons-responsive), F&R's account seems to deliver the wrong result that she is still a fully morally responsible agent.

Levy's objection using Mele's case therefore highlights how F&R's mechanism ownership condition fails to provide intuitive answers in cases featuring manipulated agents, who take responsibility for their altered values. This is only one example in a series of manipulation worries that F&R's mechanism ownership condition faces. Others include well-known manipulation worries such as Mele's (2006) case featuring the Goddess Diana and the zygote Ernie, as well as Pereboom's (2001) four-case argument. Both cases shows that F&R's mechanism

ownership condition fails to explain cases featuring global manipulation, where the agent is not, as it were, alienated from the mechanism that produces their action.[84]

I will not here attempt to defend F&R's account from these further manipulation worries, instead admitting it to be a current flaw within it. Considering also Eshleman's (2001) argument that the mechanism ownership condition is in fact superfluous to F&R's account of moral responsibility, I will forgo the mechanism ownership condition on guidance control when I attempt to use the account for autonomous systems in the following chapters. This still leaves F&R's account unable to respond to sophisticated manipulation cases. In other words, instead of mechanism ownership, I will need some kind of 'no-manipulation clause' along with moderate reasons-responsiveness in order to establish guidance control when creating a machine compatibilist account based on F&R's theory. In Chapter 6, I will come back to this problem and address it thoroughly in relation to autonomous systems.

In this chapter, then, I have explained F&R's account of moral responsibility, and in particular the key requirement of guidance control, which in turn consists of moderate reasons-responsiveness and mechanism ownership. Having shelved the issue of mechanism ownership for now, in Chapter 5 I will apply the moderate reasons-responsiveness criteria to autonomous systems, and then, in Chapter 6, I will consider how a machine reasons-responsive compatibilist account may explain worries about manipulation.

[84] By global manipulation, I here refer to a level of manipulation that is arguably analogous to agents being determined. See Pereboom (2001, 120-122) for a specific example of how the four case argument renders F&R's mechanism ownership condition insufficient for guidance control.

Chapter 5: Reasons-Responsive Machine Compatibilism

In this chapter, I will start the presentation and arguments for my reasons-responsive machine compatibilist account for analysis of autonomous systems in moral responsibility gap cases. The account itself will be modelled on Fischer and Ravizza's (1998) theory, as it was presented in Chapter 4. The full story of this machine compatibilist account will be laid out clearly over the course of this chapter and the next: this chapter will focus on the moderate reasons-responsiveness aspect and Chapter 6 will be on mechanism ownership and manipulation.

This chapter will be divided into four sections. In §I, I shall briefly set out the conditions for moral responsibility for autonomous systems based on F&R's account, which will set the tone for the sections to come. I will also clarify the notion of 'autonomy' in relation to autonomous systems, thereby making it clear which kinds of system the account will apply to.

In §II, I will argue that even some of the simplest autonomous systems may be considered reasons-reactive in so far that they are reactive to relevant inputs. I will do so by showing that the decision-making mechanisms of simple autonomous systems may be found to be strongly reactive to reasons to do otherwise in relevant nearby possible worlds.

In §III, I will further argue that simple autonomous systems can be considered regularly reasons-receptive. I will do this by showing that on F&R's account, a simple autonomous system can be found to be regularly receptive to a range of reasons granted that the mechanism is strong reasons-reactive.

In §IV, I will address the worry that autonomous systems just are not the type of entities that can be said to act/provide outputs based on reasons. To this end, I use Dennett's (1987) theory of the 'Intentional Stance' to provide an analysis of autonomous systems and their outputs.

As such, I argue that by using the 'Intentional Stance', a coherent story of how autonomous systems act/provide outputs based on reasons can be given.

By the end of this chapter, I will then have provided the first half of my machine compatibilist account based on F&R's theory of reasons-responsive compatibilism. Further, I will have shown that even the simplest form of an autonomous system can fulfil F&R's conditions for moderate reasons-responsiveness, thereby fulfilling a key component of the freedom-relevant condition for moral responsibility.

I. Autonomous Systems and Guidance Control

In this section, I will present the core claims of our machine compatibilist account using Fischer and Ravizza's reasons-responsive compatibilist theory as the foundation. These core claims will then be the focus for the latter sections. I will also make a distinction between 'inside-the-box' and 'outside-the-box' questions, and present a roadmap of the different levels of autonomy ascribed to the autonomous systems which will be discussed throughout this chapter.

In Chapter 4, I explained F&R's (1998) account of the requirements for moral responsibility: "moral responsibility – for actions, omissions and consequences – simply requires guidance control…. That is, guidance control is the freedom-relevant condition necessary and sufficient for moral responsibility" (1998, 241). Now, at a first glance, these conditions may be extended to cover autonomous systems as follows:

1. Guidance control is the necessary and sufficient condition for autonomous systems to be morally responsible.[85]
2. Guidance control consists of moderate reasons-responsiveness and mechanism ownership.
3. Moderate reasons-responsiveness and mechanism ownership are the co-sufficient and necessary conditions for autonomous systems to be morally responsible.

Of course, F&R (1998) devised these conditions with only human agents in mind. As such, an obvious reaction might be that taking these conditions and directly transferring them to the field of machine ethics is questionable. But for the sake of argument, let us begin with these conditions and work from there. I will in this chapter and the next investigate how these conditions might be manifested and attributed to different systems.

Two distinctions will be important for what follows: first, a distinction between 'inside-the-box' and 'outside-the-box' questions, and second, a distinction between different levels of autonomy that an autonomous system may possess.

First, then, the inside/outside-the-box description. Recall from Chapter 1 that the decision-making mechanism for an autonomous system may be described as a 'black box' to its creators. By 'black box', I mean that while the system is running, the decision-making process of an autonomous system is not transparent, and for most systems the lack of transparency may continue after shut-down. In other words, there is no direct access to the decision-making process of

[85] More specifically this is the criterion for fulfilling the freedom-relevant condition of moral responsibility, discussed in Chapter 2, as opposed to other conditions (such as an epistemic condition) that might also be required for moral responsibility. When discussing the applicability of moral responsibility to autonomous systems in this chapter and the next, I will be implicitly restricting myself to the issue of the applicability of the freedom-relevant condition for moral responsibility.

autonomous systems, hence after launch autonomous systems most often become a 'black box' to its creators and observers alike. For my purposes here, it will be useful to take such a description literally. When asking whether an autonomous system satisfies the freedom-relevant condition on moral responsibility, two different kinds of question reveal themselves; we can think of these as, respectively, outside-the-box and inside-the-box questions. Each aspect will illuminate different parts of the answer to the question about systems and reasons-responsiveness.

Outside-the-box questions are questions that can in principle be answered by treating the machine or mechanism as a black box, that is, by considering only its environment, input, and outputs. Inside-the-box questions, by contrast, concern the inner workings of the box. For example, one might wonder how reasons are represented and manifested internally within the mechanism. Or one might wonder whether autonomous systems can internalise reasons, more specifically moral reasons, in a relevant manner for moral responsibility.

With this distinction in place, we can see that F&R's (1998) account of reasons-responsiveness, explained in Chapter 4, is an outside-the-box account. First, they conceive of 'reasons' – the weather being poor, the drug being accessible, the queue being short and so on – in an externalist way, where 'receptivity' to reasons is a matter of the agent (or rather, the agent's action-producing mechanism) recognising these features of the external world (or of the agent itself) *as* reasons to decide or to act in one way rather than another. And second, in determining whether a mechanism is reasons-responsive, we consider what the output of the mechanism would be in a range of counterfactual scenarios. How exactly the mechanism works is not relevant; we merely need to hold fixed its working in the *same* way – whatever that is – in considering those counterfactual scenarios. More mechanistically put, the question of reasons-responsiveness boils down to: given various features of nearby possible worlds (including features of the mechanism

itself, e.g. its location and current internal state), does the mechanism detect those features, that is, have them as input, and does it then produce the appropriate output?

F&R's account was of course developed with humans in mind. As such, the outside-the-box focus makes sense, as it is assumed that all mechanisms being assessed are human in essence, and therefore roughly similar in nature: the mechanisms in question are all human *minds*. However, as we move into the realm of autonomous systems, such an assumption no longer applies: the internal mental features that humans deploy in reasons-responsive decision-making, such as beliefs, thoughts, and intentions, fall into question as the discussion turns to focus on autonomous systems. Hence, inside-the-box questions naturally arise when we ask about reasons-responsiveness as it applies to autonomous systems.

Both kinds of question, then, must be addressed in order to make an account for reasons-responsive autonomous systems. I will start by investigating possible ascriptions of reasons-responsiveness to autonomous systems by purely considering the relevant outside-the-box matters. As such, I will be discussing how the decision-making mechanism of an autonomous system might be shown to be functionally reasons-responsive based on its inputs and relevant outputs. Then in the last section of this chapter, I will address the inside-the-box questions, including addressing how one might argue that autonomous systems genuinely act 'for' reasons.

The second distinction advertised above concerns autonomy. Since the beginning of the thesis, I have used terms like 'machines' and 'systems' interchangeably, and until now there has been little need for delving into these terms further. As I am now about to consider how F&R's concepts can be used to develop a machine compatibilist account, more information is now needed.

In this chapter and the next, my focus will primarily be on autonomous systems. Machine ethics focuses primarily on autonomous mobile robots, i.e. artificial physical agents that

use sensors to observe and gather information about their environment and in turn act upon this gathered data. For ease of visualisation, I will in these chapters, like the writers I will mention below, use primarily mobile robots in my examples. Though mobile autonomous systems merely represent a sub-category of autonomous systems, they are far easier to explain and use in thought experiments or imaginable scenarios. Recall from Chapter 1 that 'autonomous' is used more liberally within the robotics and engineering discourse than it is within the field of philosophy. The word is used in this chapter in line with the ISO-Standard. More specifically, autonomy is here defined as the "ability to perform intended tasks based on current state and sensing, without human intervention" (ISO-Standard 8373:2012, 2.2).

A commonly used and slightly more specific definition is presented by Beer, Fisk and Rogers (2014): 'The extent to which a robot can sense its environment, plan based on that environment, and act upon that environment with the intent of reaching some task-specific goal (either given to or created by the robot) without external control.' (Beer et al. 2014, 77)

When the word 'autonomy', 'autonomous' or its likes are used in these chapters, they will refer to concepts in line with these definitions.

Within the machine ethics literature, there has been a tendency to tar all autonomous systems with the same brush. They tend to be envisioned as fully fledged 'agents', who decide and choose in life-or-death matters. As an example of this, one might simply look at some of the writers discussed earlier in this thesis.

Sparrow, for example, writes about AWS (Autonomous Weapon Systems): "What distinguishes AWS from existing weapons is that they have the capacity to choose their own targets. If we understand the autonomy they exercise in doing so only as a limit on our ability to predict how they will behave, then on the face of it this implies that the more autonomous they become the

less confidence we can have that they will attack the targets that we intend" (Sparrow 2007, 70). Wallach and Allen (2009), in a discussion of driverless cars, write: "Driverless systems put machines in the position of making split-second decisions that could have life or death implications" (Wallach and Allen 2009, 14). And, Coeckelbergh (2016), in his paper on the phenomenological experience of being in a driverless car, writes: "all agency is entirely transferred to the machine" (Coeckelbergh 2016, 754).

The tendency for slight overdramatization of the capabilities of autonomous systems within the machine ethics literature has not gone unnoticed. Lucas (2013) dismisses worries about autonomous systems, while slightly mocking the description of autonomous systems in the literature. He writes: "The critics appear to envision cyborgs (like 'the Terminator') or the infamous intelligent computer 'HAL' (from Arthur C. Clarke's science fiction novel 2001: A Space Odyssey) in command on the bridge of a nuclear submarine, or 'R2D2' and 'C3PO', fully weaponized and roaming the mountains of southern Afghanistan but unable to distinguish (without human supervision) between an enemy insurgent and a local shepherd" (Lucas 2013, 8).

I agree with Lucas that when worries about autonomous systems are raised, they need not be envisioned as HAL or something similarly sci-fi. I will argue that such envisioning is merely based in a misunderstanding of the capabilities of the different levels of autonomy that can be ascribed to autonomous systems.

When 'autonomous' is used to describe a system in practical engineering language, autonomy works on a scale. Adapted from the scale developed by Watson, Duecker and Groves (2020, 7), the following levels of autonomy for mobile autonomous systems can be expressed:

Level 0: No autonomy. The autonomous system is fully tele-operated by a human.[86] An example of such a system might be a child's remote-controlled toy helicopter.

Level 1: System Assistance. The system provides some automated functionality. Essentially, the system can provide assistance, but does not independently perform tasks. As an example, imagine a system making a submarine stay at a certain depth after being asked to, while the operator controls the further movement of the submarine. There is still a clear operator of the system who is actively controlling it; the system just lends assistance to the task being performed by the operator.

Level 2: Task Autonomy. At this level, the system can execute tasks independently, but does not have the ability to plan tasks or provide task strategies. The task plan and specifications come directly from the operator. An example of this type of system might be found in most car-building factories, such as welding robots. These are given a set task with clearly set parameters by the operator, e.g. 'build this section of a car', and so they perform this task over and over again.

Level 3: Conditional Autonomy. The system can generate different task strategies, but to execute one it needs human approval. As an example of a conditional autonomous system, one might imagine a surgical robot.[87] A surgical autonomous system of this level would have perceptual abilities to observe the surgical scenario and be able to plan multiple different task strategies, but would need a surgeon's approval of a

[86] Tele-operation here just meaning that the system is human-controlled remotely. Note that this level does not include what Sullins (2006) refers to as 'tele-robots', as they function with a low level of autonomy with regard to given task. Therefore Sullins' (2006) description of tele-robots would fit better on level 1 of the scale presented here.

[87] See Attanasio et al. (2021) for an explanation and analysis of surgical robots at different levels of autonomy.

specific strategy before the system can perform it. So, here the robot can create suggestions for different actions to be taken in form of surgery, the doctor signs off on it, and then the robot performs the surgery, while monitoring the environment and updating its plan in real-time.

Level 4: High Autonomy. The system can create and execute multiple tasks based on a set of boundary conditions set by the operator. The system does not require an operator's input to execute tasks, but the operator is still there to supervise. Basically, one here has a system that can create multiple types of outputs without being given human input. Only the boundary conditions are set by an operator. An example could be food delivery robots.[88] The robot will have its boundary conditions set, such as its' goal being delivering food in a given area, but otherwise left to get on with it. The robot can independently create a task strategy (e.g. to take this order first, then take this route) and execute it without needing any inputs from an operator. If operating in cities, the system needs to constantly observe and interact with its environment including interacting with ongoing traffic and so on. If operating in the countryside, it needs to adapt to rough terrain. However, if the robot gets stuck or finds itself without a task strategy, control of it may be taken remotely by a human operator if needed.

Level 5: Full Autonomy. The system requires no human inputs whatsoever. The system can be deployed into an environment and left with no human supervision. At the time of writing there is no public knowledge of a system with full autonomy. Therefore, one enters the future category. However, here is what an example of such a

[88] For a current simple example of such robots, see Rigby 2019 on Starship Technologies' delivery robots.

hypothetical system might look like. The example may be imagined in the care industry, where development of highly autonomous systems is happening at rapid speed, taking the form of a fully autonomous system in a care robot. This system would be able to observe its environment, including care home patients. It would be able to create strategies for task execution, including tasks such as the patient needing a shower, dental care or clean sheets. It would be able to execute these tasks without needing approval from a human supervisor, and update its task strategies during the execution. It would not need human supervision, but could practically just be turned on and then get to work. In short, this care robot would do its job independently, just like its human colleagues - it just might not be as great company in the break room.[89] The main difference between level 4 and 5 is the transference of control from the system to a human operator in a case of emergency or trouble. If a level 4 system gets stuck, one might expect a supervisor to be needed to help it return to its task. With a level 5 system, there is no supervision needed, as it can create task strategies to get itself back on track if stuck. The difference is therefore not in the actual presence of a supervisor, but specifically in whether the system needs supervision in certain scenarios.

Contemporary systems labelled 'autonomous' usually fit within levels 2-4 above. As mentioned, level 5 systems are not currently a reality. There are good reasons to adopt a scale of autonomy for describing the autonomous systems that I will be discussing. First, of course, we might avoid the kinds of miscommunication seen in the current literature. As such, when I develop

[89] This is of course ignoring the importance of human connection and interaction in care homes along with other ethical concerns about robotic care; however I will not get into such a discussion here. For more information about the debates surrounding robotic care, see Sharkey and Sharkey 2012, Coeckelbergh 2016 (B) and Lancaster 2019.

an account for autonomous systems and moral responsibility, it should be clear what systems exactly are encompassed by this account. Second, any objections to this account must therefore not rely on a misunderstanding of the kind of autonomous system under discussion. For example, several worries raised by writers such as Sparrow (2007) and Lin (2015) are clearly geared towards discussion of autonomous systems with the two highest levels of autonomy. Yet critics, like Stahl (2004) and Lucas (2013), rebut those worries primarily by using examples of autonomous systems with only task autonomy or conditional autonomy. By making it clear what types of autonomous systems I am writing about here, I seek to avoid these kinds of crossed wires.

Henceforth, then, I will deploy the different levels of autonomy described above for use in my machine compatibilist account, hopefully making it clear what levels of autonomous systems can fulfil the different criteria for guidance control – and which cannot.

II. Machines and Moderate Reasons-Responsiveness I – Reactivity

As we saw in Chapter 4, F&R's (1998) concept of guidance control consists of moderate reasons-responsiveness as well as mechanism ownership. Recall that F&R (1998) contend that moderate reasons-responsiveness, in turn, consists of weak reactivity and regular receptivity to reasons. In this section, I will focus on the weak reactivity criterion of moderate reasons-responsiveness and how autonomous systems might be said to fulfil this criterion.

F&R's, then, claim that moderate reasons-responsiveness merely requires weak reactivity to reasons. For our machine-focused account of moral responsibility, this condition can be reworded into our first claim about machines and reactivity to reasons:

M-Reactivity claim: To be moderate reasons-responsive, it is necessary for the mechanism in question to be weakly reactive to reasons.

In this section I will argue that even simple autonomous systems can fulfil the reasons-reactive condition of moral responsibility, when considering only the outside-the-box matters. Recall F&R's notion of weak reactivity and how it might be attributed to a potential agent. To be weakly reactive with respect to a given action, the mechanism in question must do otherwise in just one of the nearby possible worlds where there is sufficient reason to do so. Remember that we are solely focused on the outside-the-box matters here. In other words, for a mechanism to be weakly reasons-reactive in relation to a given output, the mechanism must provide a different output in a nearby possible world where there is sufficient reason to do otherwise.

To illustrate how this condition can easily be fulfilled, I will give an example. Do not be taken aback by the simplicity of the following example, but instead remember that I am only discussing reasons-reactivity here and that reasons-reactivity can be understood here to merely mean 'reactivity to relevant inputs in the form of external stimuli'. Attribution of (weak) reasons-reactivity alone does not deliver moral responsibility, and I am therefore not about to argue that even the simplest systems may be morally responsible.

Now, consider one of the earliest examples of an autonomous system. In the late 1940s, William Grey Walter designed and created small robots called 'turtles' (Wortham 2020, 36). The turtles were able to exhibit phototaxis, the simple autonomous behaviour of moving towards or away from a light source, thereby falling within level 2 of the autonomy scale presented in §I. Yet phototaxis, in all its simplicity, gave rise to fairly complex movements, and Walter's turtles caused much amazement because the turtles' movements made them seem 'alive' to a 1940s/50s audience. I will use Walter's turtles in my example here, because they represent the most basic of autonomous

robotic design being the better part of a century old. As such, any conclusions about a turtle's reactivity or receptivity to input can be assumed to be true of much more complex modern systems as well.

So, imagine that one of Walters' turtles is placed in the middle of a sizeable room. As the turtle is placed, a light is turned on in the left side of the room. This turtle has been designed to move towards light, and so, as it exhibits phototaxis, it moves towards the light.

Before taking a closer look at this example, one thing must be addressed, namely the relevant mechanism in this case, namely the turtle's decision-making algorithm. For a turtle, this algorithm is simple as it is only designed to interact with one kind of external stimuli, namely light. In layman's terms, the turtle has been given a goal by its operator in the form of a command 'go towards the light'. If there is a light source, the mechanism will make the turtle move towards it. If there is none, the turtle will stand still. Furthermore, if there were a second, competitor source of light, we might imagine the mechanism to guide the turtle towards the stronger source. In this way, the turtle is able to execute the task without intervention from a human operator.

The question at hand then, is: is the turtle in the above example weakly reactive vis-à-vis its movement to the left? To answer this, I will analyse the mechanism of the system in question in the same manner as one would do with a mechanism belonging to a human agent. In short, I shall consider the mechanism as a black box and consider the factors external to this box. In the described scenario, the question is whether the mechanism is weakly reasons-reactive regarding the decision to move to the left.

To find out, we may consider nearby possible worlds, where the mechanism operates and there is sufficient external reason for it to not move to the left. Imagine a nearby possible world, where instead the light source is placed on the right. The placement of the light source at the right

can sensibly be taken as a sufficient external reason for moving towards the right for any agent whose goal is to move towards the light. Then, when holding the relevant mechanism fixed, the turtle will indeed move to the right when presented with sufficient external reason to do so in this nearby possible world.

Based on the outside-the-box factors, then – or, in other words, when considering purely the input (the external reason) and output (the described movement) of the mechanism in a relevant nearby possible world – the light-sensitive mechanism of the turtle is weakly reasons-reactive in relation to its output in the actual world.

To push this conclusion further, I will argue that autonomous systems of a wide variety and including something so simple as turtles may even be found to be not merely weakly but strongly reasons-reactive. Consider the turtle case again. Plausibly, in *all* the nearby possible worlds where the relevant mechanism is held fixed and the light source changes its place to somewhere else in the room, the turtle will move towards the light. Therefore, if in a nearby possible world, the light source was in a south direction, then south the turtle would go. The same may be found with any other direction. So, the turtle would seem to be not just weakly, but indeed strongly reasons-reactive.

One might object that cases could be imagined where the turtle would fail to react appropriately moving towards the light. For example, one could imagine a possible world where there is a malfunction of some kind, so that the source of light is registered, yet the mechanism fails to provide the expected output; or where a lab student alters the turtle's mechanism so that it will move towards the left independent of the position of nearby light sources. However, in any such possible scenarios, the relevant mechanism in the original case cannot be said to have been held fixed. Thus, such scenarios are irrelevant to the mechanism's degree of reactivity. Moreover, a

system such as a turtle does not and cannot suffer from weakness of the will or similar human phenomena that could render it unresponsive to relevant inputs. If no physical restraints are put on the system and the mechanism is truly held fixed, then the mechanism will react to the external reasons for movement presented in form of light in any given nearby possible world. Hence the mechanism of a system as simple as a turtle can be strongly reasons-reactive. Clearly then, even simple systems can and often do satisfy the 'weakly reasons-reactive' part of the requirements for moderate reasons-responsiveness.[90] Now, remember that I am here only saying that simple autonomous systems may be considered reasons-reactive in so far that they are strongly reactive to relevant inputs.

There are, however, some potential objections to the claim that the turtle is (at least) weakly reasons-reactive. I will briefly address three such objections here. The first objection concerns the use of the term 'reason' in relation to the autonomous systems mentioned in this chapter. Specifically, one might object that the turtle does not, strictly speaking, have any *reasons* for acting at all; it has merely been programmed to move towards emissions of light. However, to queries on that topic, I must reply with a reminder of the distinction between outside-the-box and inside-the-box matters. As I explained in §I above, F&R's account makes use of an externalist conception of 'reasons' to analyse the reasons-responsiveness of human agents' mechanisms. As such, the continued use of that externalist conception in the machine compatibilist version of the account can hardly surprise the reader. Hence, if one has something to say against the use of 'reasons' in this sense in this section, it must instead be based in scepticism about whether the system can internalise reasons for action – for example, whether the system can represent (external)

[90] This point might be of particular interest to those, who find weak reasons-reactivity to be inadequate for being a co-sufficient condition of moral responsibility in F&R's original account. See Mele 2000 and 2006 for a developed discussion of this objection. See Fischer and Ravizza 2000 for their initial response.

reasons *as* reasons, and/or whether it has the ability to act 'for' reasons. Such matters are inside-the-box questions, which I am deferring until the last section of this chapter.

The second objection concerns the simplicity of the system used in the example. One might suggest that ascribing reasons-reactivity to such a simple thing is mere hogwash. After all, our capable turtles cannot react to any other factors other than light – they do not have human-like complexity in their decision-making and reactivity.

I will easily contend that turtles and similar simple systems are, of course, not complex or nuanced decision-makers. Asking them to solve Sophie's choice will not do us any good. I do not wish to somehow mislead anyone to think these machines to be anything more than they are. However, such nuance or complexity is not required for reasons-reactivity, as defined by F&R. Indeed, notably, the fact that the turtle can only recognise one kind of reason (emission of light) and react in a particularly limited way (movements towards it, if light is present) is irrelevant to the ascription of reasons-reactivity.

First, the fact that the turtle can only recognise one kind of reason is a matter of receptivity – the topic of the next section – and not reactivity. Second, reasons-responsiveness is defined by F&R with respect to a single, specific action of the agent. For example, if we wish to determine whether I am reasons-reactive with respect to my action of guiding my car to the left, then my abilities or tendencies to dance, write or yell play no part in this assessment. In other words, the action being questioned is assessed separately from the agent's other doings and capabilities. Hence, when determining whether the turtle is reasons-reactive in relation to its movement to the left, its simplicity as a system makes no difference to the assessment. More specifically, its simplicity is not a hindrance when treating the mechanism as a black box and assessing its reactivity merely in relation to those external factors to which it is receptive.

If the reader is still sceptical, consider this analogous case featuring a human agent. Imagine a woman who is lost in an underground complex of caves and looking for a way out. She sees a small glimmer of light in the far left and as such goes towards it. The relevant mechanism here may be supposed to be the woman's faculty for rational decision-making – a faculty that, thanks to the woman's sensory equipment, is sensitive to the presence of a light source, and which, in the circumstances, leads her to move towards any light source because she sensibly believes that the light will be daylight, and hence indicates a way out of the caves.

With respect to her action of moving to the left, we may suppose that the woman is at least weakly reasons-reactive. As such, suppose in nearby possible worlds, where the light emits from various other directions, the woman's decision-making mechanism is reactive to the light as a reason to move towards it. The woman and the turtle thus exhibit analogous behaviour in their respective world and other relevant nearby possible worlds, and are thus – when it comes to their reactivity to reasons (this being the emissions of light) to move in one direction or another – in the same boat.

When assessing the reasons-reactivity of the two mechanisms, different as they are on the inside, their inputs and outputs are the same. Of course, the woman can easily be assumed to have a rich mental life and her actions might be fuelled by a sophisticated human understanding of, for example, why the presence of a light source constitutes a sufficient reason for her to move towards it. However, and I apologise for sounding like a broken record, objections relating to a potential need for such sophisticated understanding are inside-the-box matters and can therefore be put aside for now. If we treat both the woman and the turtle as black boxes and purely look at what goes in and what comes out, they are – in this specific and very limited situation – the same.

The last objection that I will discuss here concerns reactivity to specifically moral reasons. One might object that through using turtles as examples, I have neatly avoided addressing the topic of systems' reactivity to moral reasons. My answer to this worry here will be brief. What must be remembered when considering the reasons-reactivity of turtles and the likes is that only the reactivity to external factors is being assessed. Recall that in Chapter 4, I explained how an agent could be strongly reasons-reactive while only regularly reasons-receptive. A mechanism's degree of reactivity is to be measured only in relation to those reasons to which it is receptive. So in cases, such as the turtle, where the mechanism is only receptive to a very narrow range of reasons, this is no bar to its being reactive *to those reasons.*

Autonomous systems, as I have shown, can be reactive – indeed, strongly reactive – to reasons, externalistically understood; as with the turtles, these usually take the form of sensory inputs. Of course, further questions naturally arise about what types of reasons a system can be receptive to, and in particular whether autonomous systems can ever be receptive to the types of moral reasons that are usually relevant to moral responsibility. Worries about whether systems of any kind can be receptive to moral reasons is a valid concern, but – as I have said – it does not affect the question of reactivity to reasons. I will therefore address this concern – which is a concern about receptivity rather than reactivity – in the next section, where I discuss autonomous systems and reasons-receptivity.

In this section, I have shown that even the simplest form of level 2 autonomous systems (turtles) can be found to be strongly reasons-reactive, thereby in theory fulfilling the reactivity element of moderate reasons-responsiveness. Hence, if one wishes to deny the possibility of autonomous systems having guidance control, one must look elsewhere for grounds for doing so.

III. Machines and Moderate Reasons-Responsiveness II – Receptivity

In this section, I will argue that autonomous systems can be found to be reasons-receptive regarding their output in certain scenarios. By the end of this section, I will then have shown how autonomous systems can, in the given circumstances, be moderately reasons-responsive.

Recall from Chapter 4 that F&R (1998) conclude that moderate reasons-responsiveness requires regular receptivity to reasons. This requirement, expressed for the machine compatibilist account, may then be summarised as:

> M-Receptivity claim: To be moderately reasons-responsive, it is necessary for a machine to be regularly receptive to reasons.

I will argue that autonomous systems can fulfil this criterion. To do so, I will start by leading us through some examples and then move onto a discussion of different types of reasons. First, recall F&R's specific definition of regular receptivity: 'Regular receptivity to reasons, then, requires an understandable pattern of reasons-recognition, minimally grounded in reality' (F&R 1998, 73).

Recall from Chapter 4 how regular reasons-receptivity was attributed, as was shown using the case of Charis buying an ice-cream cone on a nice day at the park. In the actual sequence of events, Charis recognises the sufficient reasons present for buying an ice-cream cone, i.e. the sun shining, the lack of queue and so on – and so she acts upon this recognition of reasons. However, to show that Charis is regularly reasons-receptive in this case, Charis has to exhibit a sensible pattern of reasons-recognition in a series of nearby possible worlds where she has sufficient reason to do otherwise. As was shown in this case back in Chapter 4, Charis recognised the length of the queue as a sufficient reason to do otherwise in the nearby possible worlds, in which the queue was respectively 10, 20, 30 metres and so on. According to F&R's theory this constitutes a sufficient

exhibition of sensible reasons-recognition to attribute regular reasons-receptivity to the given mechanism. With queue lengths, how a sensible pattern of reasons-recognition takes form is fairly obvious: if a queue length of X is recognised as a reason to do otherwise, then sensibly any queue length greater than X should also be recognised as a reason to do otherwise. This is what Charis exhibits in the nearby possible worlds where these queue lengths are relevant, thereby displaying an understandable pattern of reasons-recognition.

If I was about to argue that a normal human agent was regular reasons-receptive, then this would be a fairly simple task, as we in practice think most human agents, at least most of the time, are regularly reasons-receptive with respect to their actions. When uncertainty creeps in, when dealing with humans, we can sometimes in principle find out by asking them about their decision-making and their recognised reasons for action. Even when that is not an option, we usually have so much experience in how the human mind works in practice that reasonable interpretations of others' reasons-receptivity can be made. As an example, a statement like the following is fairly common: 'I do not think that she did not get you a present to be mean – to be honest it probably never even crossed her mind what date it was! If she had been aware of the date, she would have gotten you a present'. As such, it is not uncommon for humans to make educated guesses about other agent's reasons-recognition and even consider whether they would have recognised sufficient reasons to do otherwise had the circumstances been different. Yet, when it comes to attributing regular reasons-receptivity to autonomous systems, it becomes more complicated, as we cannot make assumptions about the system's reasons-receptivity based on a familiarity with the system's mental make-up, as one has tendency to do when imagining other human agents.

I will argue that functioning autonomous systems can nonetheless be found to be regular reasons-receptive while referring solely to outside-the-box matters. Consider again the

scenario featuring the phototaxis-displaying turtle from the previous section. Recall that the turtle was strongly reasons-reactive vis-à-vis light input, such that in the actual sequence of events it moves to the left towards the light, yet we have good reason to think that in nearby possible worlds where the light is in another direction (holding the turtle's light-following mechanism fixed), the turtle will move in that given direction. In the previous section, I concluded from this finding that the turtle was strongly reasons-reactive in respect to its movement towards the light.

How may we infer from this that the turtle is regularly reasons-receptive? The answer turns on the fact that reasons-reactivity is, as F&R put it, 'the capacity to translate reasons into choices (and subsequent behaviour' (1998, 69). We know that our turtle in question is strongly reasons-reactive in respect to its movement to the light. In virtue of that, we also know that the turtle is at least regularly receptive to the position of the light source as well, since if it was not receptive to that, the turtle would not be able to react in the way that it does. With well-functioning autonomous systems (and robots in general), we know that outputs require specific inputs. As such, without receptivity to the given input (e.g. the light), we would not observe the corresponding output (e.g. the movement towards the light). Hence, if a strongly reasons-reactive autonomous system exhibits a regular pattern of reactivity in a series of nearby possible worlds, then we have good reason to think that it is also regularly reasons-receptive.

However, tying attribution of reasons-reactivity and reasons-receptivity together like that has its own consequences. If our turtle fails to be reasons-responsive – if, for example, were the light source to be moved around the room, and the turtle just stood still or started moving in random directions – then we cannot tell whether this is due to a failure of receptivity or of reactivity. It could be a failure of receptivity, if the sensors did not work properly in the given scenario, or it could be a failure of reactivity, if the system failed to be reactive to a very specific input, such as

light coming in from a specific angle. However, if the system is working fine in a given scenario, then it's a fair bet to say that the turtle is both strongly reasons-reactive and at least regularly reasons-receptive.

Of course, the turtle's reasons-reactivity and -receptivity depends on what the turtle would do in a range of counterfactual scenarios, and our conviction that the turtle can be ascribed these would commonly in practice depend on observing the turtle's actual behaviour across a range of actual situations. Such evidence is not conclusive; in particular, the turtle could in principle seem to be reasons-responsive because it moves towards the light in all of our trials, yet in fact it might be behaving randomly and by sheer luck the random behaviour exhibited matches a sensible pattern. Nevertheless, it would still be true that insofar as our evidence and knowledge of the system leads us to conclude that the turtle is reasons-reactive, it equally leads us to say that the system is regularly reasons-receptive, since one must recognise a reason in order to be able to react to it.

Markedly, the same point may be made about human agents and reasons-responsiveness as well. When ascribing moderate reasons-responsiveness to human agents in practice, we cannot do it with complete certainty, since it requires facts about how the agent would behave in a range of counterfactual scenarios, and our evidence concerning those facts can only ever be indirect and hence is fallible, since by definition counterfactual scenarios do not actually happen. Furthermore, in cases where a human agent fails to do what there is sufficient reason to do, often further investigation will be required in order to ascertain whether their failure was a failure to recognise reasons or, instead, a failure to react to (recognised) reasons. As an example, if my friend did not get me a birthday present, then without further investigation I do not know if its due to my friend simply not realising the date or whether my friend did know it was my birthday and did recognise this fact as a sufficient reason to buy me a gift, yet still failed to do so. Again, however,

insofar as we have good reasons to attribute strong reasons-reactivity to a human agent, we generally thereby also have good reasons to attribute regular reasons-receptivity to them.

In effect, what I have shown so far is just that properly-functioning autonomous systems will generally be regularly receptive to the reasons that cause their outputs. However, concluding on that basis that autonomous systems are often regularly receptive to reasons *simpliciter* – and hence moderately reasons-responsive, as I argued in the previous section that autonomous systems can be, and often are, strongly reasons-reactive – would be poor form. Recall the objection that was left hanging at the end of the previous section. Our turtle is, I argued, strongly reasons-reactive in the sense that it reacts appropriately to all the reasons it is receptive to. Yet, of course, the turtle is receptive to just a single kind of reason, namely the existence and placement of a light source. That is the only kind of input that the turtle is capable of receiving and responding to. One might object that displaying (across the relevant counterfactual scenarios) an understandable pattern of recognition of such a narrow range of reasons does not really constitute 'regular reasons-receptivity'. In particular, since regular reasons-receptivity is supposed to be relevant to the attribution of moral responsibility, regular reasons-receptivity must include receptivity to *moral* reasons. We are now in a position to address that objection.

Many actual and near-future autonomous systems are, of course, considerably more sophisticated than turtles, yet they generally are still created to be sensitive to a specific, and rather limited, range of inputs. In some respects, such systems are often much more sensitive to their specific range of inputs than human agents, for example sensitivity to temperatures, precise distances, or other such data inputs. Nevertheless, they do not grapple with moral dilemmas in the sense that human agents often do. They do not ascribe abstract moral values to their data and apply it when they provide outputs, or at least they do not ascribe anything that the system itself

recognises or understands as moral values. There are projects happening designed to make systems act morally, and maybe even to take moral considerations into their decision-making process.[91] However, it is genuine moral understanding and receptivity to moral reasons that is being questioned here and it is this understanding that autonomous systems lack. The objection then takes form: if autonomous systems are not receptive to moral inputs, how can they be reasons-responsive in a way that is relevant to moral responsibility?

To answer this, I will suggest there are actually two different key questions in play in this objection, and that only one of these questions is relevant to the ascription of F&R's moderate reasons-responsiveness. Let us then start by breaking the objection into two separate questions and deal with them in turn:

1. Can autonomous systems be regularly reasons-receptive qua F&R's definition without receptivity to moral reasons?
2. Can autonomous systems be ascribed moral responsibility without being receptive to moral reasons?

Recall, both from earlier in this section and from Chapter 4, that regular reasons-receptivity merely required the mechanism in question to exhibit a sensible pattern of reasons-recognition in a series of nearby possible worlds where there is sufficient reason to do otherwise. Hence, in Chapter 4, it was possible to ascribe regular reasons-receptivity to our agent, Charis, purely based on her sensible pattern of recognising different lengths of queues for the ice-cream van as sufficient reasons to do otherwise. As such, the ascription of regular reasons-receptivity was done based on receptivity to just a single dimension of reasons, namely the queue lengths. Ascription of regular reasons-

[91] See Allen et al. 2000, Wallach et al. 2008, as well as Wallach and Allen 2010 for more on this type of project.

receptivity on F&R's account is permissible, even if the agent's receptivity to reasons only covers a specific range of reasons. As such, regular reasons-receptivity was also ascribed to our turtle earlier in this section, despite only being sensitive to light input. F&R's concept of moderate reasons-responsiveness applies with respect to a given action, e.g. to determine whether Charis was reasons-responsive with respect to her purchase of an ice-cream cone. The attribution of reasons-reactivity or reasons-receptivity to a given mechanism is therefore not limited to scenarios featuring moral reasons, nor does it require the mechanism in question to be sensitive to such reasons. In short, to answer question 1: yes, autonomous systems can be regularly reasons-receptive without being receptive to moral reasons. To put it another way, recall that moderate reasons-responsiveness is F&R's account of agents' 'guidance control' with respect to specific actions, and such control need not require receptivity to moral reasons. After all, the example F&R use to introduce guidance control in the first place – where Sally has guidance control with respect to turning her car – is not one where moral reasons for action are in play.

That said, this thesis is about moral responsibility gap problems, and so far, I have been discussing how one might open up the conversation surrounding these problems by focusing on autonomous systems and how they might fulfil the freedom-relevant condition for moral responsibility, namely guidance control. Hence, question 2 is a sensible question to ask: after all, if autonomous systems are only sensitive to non-moral inputs, thereby making them unlikely candidates for moral responsibility, what does their potential reasons-responsiveness matter?

To answer this question, I must ask the reader to remember the overarching goal of this thesis, as discussed in the introduction to Chapter 3. My aim is to shed some new light on moral responsibility gap problems by developing a machine compatibilist account founded on F&R's reasons-responsive compatibilism. As mentioned previously in both Chapter 2 and 3, my goal is

therefore not to conclude that we can, in fact, ascribe moral responsibility to autonomous systems, but instead to see what happens to our understanding of moral responsibility gap problems when the systems are analysed as entities that might fulfil some of the conditions for moral responsibility. This is an important distinction. I have in this chapter shown that autonomous systems may be considered moderately reasons-responsive. This is significant in the context of moral responsibility gap problems because, as we saw in Chapter 1, the literature on that problem persists in seeing autonomous systems as just very sophisticated tools. Seeing them in that light closes off the possibility of even sensibly *asking* the question whether they are or could be themselves bearers of moral responsibility: asking whether something that is the mere tool of a human agent might be morally responsible for anything is absurd.

Yet, I have argued in this chapter that autonomous systems differ greatly from mere tools, and that difference is, precisely, that they can be – and are – responsive to at least some reasons. Hence, they are, in an important sense, in control of their own behaviour/outputs. That makes the question of their potential for moral responsibility a sensible one to ask, rather than an absurd one. I have not suggested, nor will I suggest, that the moderate reasons-responsiveness that can be attributed to autonomous systems is somehow sufficient for moral responsibility.

In Chapter 6, I will argue that autonomous systems, albeit they can be moderately reasons-responsive, nevertheless fail to fulfil the full conditions for guidance control and thereby the freedom-relevant condition for moral responsibility. More specifically, I will argue that current and near-future autonomous systems fail to satisfy the aspect of guidance control that requires an agent to not be manipulated. One might then further – or instead – argue that sensitivity to distinctively moral reasons is required for moral responsibility, and that autonomous systems (at least current ones) fail to be morally responsible for that reason. Even so, the argument that

autonomous systems are moderately reasons-responsive will nonetheless have achieved its aim, since to claim that moderate responsiveness to reasons that include distinctively moral reasons is required for moral responsibility is to grant my main point: that by satisfying *a* necessary condition on moral responsibility, namely moderate reasons-responsiveness more generally, autonomous systems are not to be conceived as mere tools of human agents, and hence that we need to consider more carefully exactly what requirement on moral responsibility they are failing to meet. This is a question that goes unasked if we insist on conceiving of autonomous systems as mere tools.

It should now be clear, then, how reasons-responsiveness matters, even in the absence of receptivity to moral reasons and even without (yet) an answer to the question of what further requirements there are on the attribution of moral responsibility. Having shown that autonomous systems fulfil the conditions for moderate reasons-responsiveness, it opens up for new perspectives when analysing moral responsibility gaps, as we no longer deal with autonomous systems as mere tools, but instead as reasons-responsive entities. I will in Chapter 6 go into more detail about how the attribution of moderate reasons-responsiveness alters the way we should think about moral responsibility gap problems.

IV. Robots, Reasons and the Intentional Stance

In the two previous sections of this chapter, I argued that autonomous systems can fulfil the conditions for F&R's notion of reasons-responsiveness without reference to their internal makeup. However, to the sceptic's eye, fulfilling these counter-factual conditions is only half the battle, as the most bothersome leftover question is still in play: are autonomous systems the kind of things that can be genuinely said to act for reasons? In this section, I will address this query.

In the previous sections and the previous chapter, I explained how F&R's conditions for reasons-responsiveness constitute a modal constraint on the inputs and outputs of a given mechanism. In §I, I summarised this as the idea that reasons-responsiveness is an 'outside-the-box' matter – the metaphorical box here being the vessel of whatever mechanism is being assessed. Reasons-responsiveness can thus be attributed to an agent or thing, including an autonomous system, without requiring it to be conscious, or to have a mind, or similar. We need not worry about the 'ghost in the machine' in order to show that an autonomous system is moderately reasons-responsive.

For humans, we of course have a general understanding of what is the 'inside the box' for ourselves. Reflection upon our own beliefs, desires and reasons all come as part of our mental makeup. In §III on reasons-receptivity, I argued that autonomous systems could be sensitive to reasons in a manner sufficient for moderate reasons-responsiveness. However, a sceptic might wonder if proven sensitivity to reasons in this sense – that is, sensitivity to features of its environment to which it needs to respond if it is to reach its goals – is enough to really show autonomous systems to be capable of reasons-responsiveness. F&R's account makes it clear that it is not just sensitivity to reasons, including moral reasons, which fuels reasons-responsiveness. When considering the acceptable patterns of reasons-receptivity, F&R write: "the pattern in question must show that the agent (when acting on the actual mechanism) recognizes that the other person's claims give rise to moral reasons *that apply to him*" (1998, 77). This seems to suggest that reasons-receptivity requires the ability to *reflect* upon one's beliefs and the needs of other people, as well as recognising how these give rise to moral reasons.

F&R go on to clarify this further, and the focus is clear: what is seemingly needed is for the agent to be able to reflect and feel the weight of the normative constraints that arise from

being in a moral community. F&R argue that this reflective aspect of reasons-responsiveness excludes a range of creatures and potential agents from consideration. For example, they write: 'One stock example is that of intelligent animals, like dolphins and higher-order primates. Although such creatures appear to act on mechanisms that respond to a range of incentives, it is highly doubtful that they have any adequate grasp of notions like moral rights and duties (or moral reasons in general)' (F&R 1998, 78). It is this 'grasp' of these concepts that one may naturally doubt that autonomous systems are able to have – indeed the sceptic can easily and fairly assume that no system has been created that has shown such human-like reflection when it comes to these matters.

As I showed in the previous section, autonomous systems' ability to be sensitive to external factors and create appropriate outputs in response is not in doubt. Instead, what the sceptic must question here is the possible system's reflection and understanding of the reasons upon which it acts. As such, the full potential objection that lies here must be something along the following lines: it has been shown that autonomous systems can in theory be sensitive to the right kind of external factors and in turn create appropriate and 'rational' outputs, which seems to fulfil the conditions for reasons-responsiveness. However, autonomous systems cannot be shown to be reflective or have 'human-like' understanding of the reasons upon which they act. They are not the kind of things that can reflect upon beliefs about themselves and the people around them nor clearly can they make moral demands on themselves based on such reflection.[92] Without a grasp of the reasons in question, autonomous systems cannot be the kind of thing that may be considered reasons-responsive.

[92] This proposed objection is inspired by the accounts of Brey 2013 and Purves et al. 2015, who both argue that robots are incapable of agency due to a lack of acting based on respectively beliefs/desires and reasons.

As such, it looks like the 'inside-the-box' questions, which I parked earlier, have finally come back to haunt us. I shall therefore here try to provide an answer to the 'inside-the-box' question: how can we attribute understanding of reasons to something that is seemingly nothing more than data, sensors, wires and flashing lights?

Dennett (1978) can help us in this endeavour. In this section I will seek to adapt Dennett's 'intentional stance' for use in explaining the outputs of autonomous systems in terms of reasons-receptivity. To start, Dennett argues that when explaining or attempting to predict something's behaviour, one might do so from one of three stances. I shall here summarise these in relation to machines.

First, we find the physical stance. Using the physical stance, we may explain or predict a system's outputs by appealing to the system's actual physical state and the relevant natural laws. Machines can reasonably be assumed to be, at the time of writing, completely determined systems. Using this stance is, in effect, doing the practical legwork of using deterministic laws to predict or explain behaviour. This may sound rather formal, but it is quite simple. Take a computer, like the laptop I am working on right now. When I give the laptop a certain input, such as hitting the power button, or clicking the save icon in Word, one can point to the computer's programming in order to predict the relevant output, here being turning on/off the laptop or saving a document. As such, one can point to that same programming, together with the relevant input, in order to explain the given output. The laptop's programming isn't corrupt, I have not spilled my tea all over the keyboard, ergo the laptop is not malfunctioning or damaged and as such we can point to the programming of the laptop, its wiring and its algorithms to predict or explain its outputs.

Second, there is the design stance. When using the design stance, one explains and predicts the output of a given system by thinking of the system as having a specific design or

purpose. As an example, consider a robotic vacuum cleaner going about its job in a flat. Where is the Roomba going next? To answer this, it is not sensible to start looking at the question from the physical stance. I do not begin some long string of calculations involving the Roomba's sensor data and the laws of its programming – even though if I did have enough information and enough time to perform the calculations, I would indeed be able to predict its behaviour. No, answering the question is much more straight-forward than that. Suppose I can see that the Roomba has yet to vacuum the rug next to it, and I know the Roomba is designed to clean in an efficient manner. In that case, I may reasonably make the prediction that the Roomba will clean the rug next based on my knowledge of its design. Explaining the Roomba's output is similarly easy. Why did the Roomba clean the rug? Well, cleaning the rug is its set task and it had yet to do it: that is why.

Of course, using the design stance in the Roomba example will seem obvious to most readers, as most people encountering a Roomba might know its design but have no chance of knowing the intricacies of the system's programming. However, Dennett is clear that the design stance is not merely to be used in lieu of the physical stance in cases of ignorance; rather, the design stance generally makes for a more apt and useful explanation of certain behaviours than does the physical stance. Even if one did know the system's programming and its sensorial inputs in the specific scenario, the explanation the physical stance would produce would simply not be as good as the one provided by the design stance. Thus, even if multiple stances can be used to explain the same scenario, one stance will provide a more fitting picture – a *better* explanation – of what is going on. Hence, in the case of a non-malfunctioning designed everyday systems, and even autonomous systems of autonomy levels below 2, the design stance will generally be the appropriate stance to adopt.

Last, but not least, one finds the intentional stance. When using the intentional stance, one explains or predicts some output or behaviour using such concepts as beliefs, goals, and intentions. Dennett summarises the use of the intentional stance for systems as follows: "one is viewing the computer as an intentional system. One predicts behaviour in such a case by ascribing to the system *the possession of certain information* and supposing it to be *directed by certain goals*, and then by working out the most reasonable or appropriate action on the basis of these ascriptions and suppositions" (1978, 6-7). In other words, in using the intentional stance, one attributes certain informational states to the agent or thing in question. One explains and predicts something's behaviour based on ascriptions of goals, intentions, and/or desires to bring about certain states.

It is of the utmost importance to note that viewing a system as intentional in Dennett's sense does not require using worrisome concepts such as consciousness, personhood or anything similar. The intentional stance is merely a route for an observer to explain and predict a system's behaviour. This is not to say that the intentional stance should be confused with the design stance. To see the contrast and difference between the two stances, consider again the Roomba. As it turns to vacuum the rug, we may explain its output through reference to its design. It is designed to do a job, here vacuuming the rug, and this stands as the explanation when it starts zooming around. However, this stands in stark contrast to saying that the Roomba has vacuumed the rug because it wanted to or because it intended to. Indeed, an application of the intentional stance is completely unnecessary: it does not provide a clearer or more helpful understanding of the Roomba's output than the design stance. Therefore, the intentional stance is not a catch-all for every type of machine or robot out there. The difference between the design and the intentional stance will become clearer in a moment when I will be discussing them further in relation to autonomous systems.

In this section I will argue that Dennett's intentional stance can in principle be used to ascribe intentional states, and in turn rationality, to autonomous systems. In short, I will in this thesis use Dennett's intentional stance as the explanation for the dreaded 'inside-the-box' matters in relation to autonomous systems and reasons-responsiveness.

To start, let us consider what happens, when one applies the intentional stance to something successfully. What does this ascription of beliefs and desires have as a consequence? For an answer, consider the following claim by Dennett: "*What it is* to be a true believer [a genuine agent] is to be an *intentional system*" (Dennett 1987, 15).[93] A common interpretation of Dennett is merely to take this quote and its implications at face value.[94] This makes the intentional stance the last frontier, when it comes to figuring out whether something has 'real' beliefs and desires. Ascribing beliefs, intentions, and so on to an intentional system is as real as it gets, or in other words all there is to being an intentional system or having intentions or having beliefs is just the fact that the stance is the most appropriate one for an observer to take in order to explain and predict the system's behaviour. Further questions about whether these concepts 'really' apply to the systems are nullified.

Though controversial, I will be assuming the truth of Dennett's account of the intentional stance, interpreted as just described.[95] I will defend this move later in this section and also, discuss other potential interpretations, however for now, one might see it as a move made to get the conversation going. In particular, I shall argue that to explain or attempt to predict the outputs of an autonomous system, we do *in fact* typically adopt the intentional stance towards non-

[93] The equalising of 'true believer' with 'genuine agent' is from Elton's (2003) work on Dennett.

[94] See Elton 2003 and Shoemaker 1990 for examples and discussion of this interpretation.

[95] The controversy stems from the account not being realist enough to satisfy some critics. See Eronen 2017, McCulloch 1990, Pöyhönen 2014, Rey 1994 and Slors 2007 for examples. See Dennett 1991 and 2000 for a response to this criticism.

malfunctioning autonomous systems; and this provides evidence that the intentional stance is a common-sense stance to adopt when dealing with such systems.

To do so, I will here consider a particular case – Google DeepMind's AlphaGo, a Go-playing system created using deep neural networks and machine learning[96] – using the different stances. My use of the case is similar to Dennett's use of an example featuring a sophisticated chess machine (Dennett 1978 4-7). However, I have chosen AlphaGo because Go is a strategic game that famously far surpasses games such as chess in complexity. As such, the use and appropriateness of the intentional stance should become even more apparent than in Dennett's (1978) original (and now rather old) example.

Now, let us imagine that I am in the middle of a game of Go against the greatest player in the world, AlphaGo.[97] Now for appearances sake, let us also imagine that I am a habile Go player. I am currently trying to predict what move AlphaGo might make next.

To start, suppose I attempt to use the physical stance to predict its move. After all, the system is determined, so the physical conditions of the system along with knowledge of the laws of its programming can yield an answer. It is obvious that taking this stance is impossible in practice. The calculations are herculean at best, and though I think myself nifty with numbers, this is well above my paygrade. The variables and minutiae of the calculations, as well as the vast number of strategic possibilities in a Go game, would break even the finest computer.[98] A system being determined does not entail or necessitate the practical possibility of human knowledge or understanding of how to predict its outputs. Moreover, even if I somehow could acquire the

[96] For more information about AlphaGo, see Silver et al. 2017 and Silver et al. 2018.

[97] AlphaGo shot to fame after beating the previous world champion of Go 5-0.

[98] It might be worth pointing out here that Go famously has more possible board configurations than there are atoms in our universe.

knowledge required to predict AlphaGo's next move using the physical stance, the sheer amount of time and complexity involved would mean that use of this stance would only be a matter of last resort, if no other route to prediction was available. Further, here in the real world I am in the middle of the game, and don't have nearly enough time at my disposal to make the calculation, even if in principle I could. Matters are even clearer when it comes to the retrospective explanation of AlphaGo's moves: reams of vast and complex calculations generally do not make for useful, comprehensible explanations. Hence, the physical stance cannot help in my grand match against AlphaGo.

There is a potential use for the physical stance that is worth mentioning though, namely in the case of malfunctions. Suppose that I am playing AlphaGo and I look at his screen, which has bluescreened. Now, in the case of a malfunction, we might use the physical stance to explain AlphaGo's output. Why did AlphaGo bluescreen? Because there was a loose wire or a software update issue or something akin to that. Similarly, people also sometimes revert to the physical stance with human agents when they act sufficiently irrationally or display strange patterns of behaviour. This casual practice is finely summarised in the colloquialism: 'they must have a screw loose!'.

Suppose that I try instead to make a prediction using the design stance. Dennett says: 'The essential feature of the design stance is that we make predictions solely from the knowledge or assumptions about the system's functional design, irrespective of the physical constitution or condition of the innards of the particular object' (1978, 4).

So, I sit here in my chair facing my opponent and I know AlphaGo's function based on its design: namely to play Go.[99] I also know that it uses neural networks for its decision-making. Of course, the problem with facing a system that uses neural networks means that their decision-making is a black box even to its designers (Matthias 2004, 178). Using the design stance does get me much further than the physical stance. Using the design stance, I can know that AlphaGo plays Go, has played a great deal of Go matches and according to its design it should have learned from these. I can further know that AlphaGo will not make any illegal moves, or concede defeat when it is in a winnable position. I can try to use the basic question of: 'What is the best move here for my opponent?', however Go is famously a game of human intuition and strategy – there is no singular 'best' move. So, I might make the assumption that AlphaGo will not make any obviously bad moves as it can be expected to have learned how to avoid these during its learning, but not much more information. Any assumptions I might make about it having strategies or similar is not reflected in the system's design. It was never 'fed' a strategy book, like one might think would be done with a chess computer, and the quality, extent and specific aspects of its learning is unknown, when using the design stance. As such, I know that it has played Go over and over again. Further than that, it is a black box in terms of design.

The trouble with using the design stance on most autonomous systems compared to standard machinery or robots stems largely from the learning component of these systems. As such, a large part of the behaviour of learning autonomous systems is not explicitly built into their design. Therefore, any explanation of AlphaGo's moves using the design stance will be incredibly superficial, as its learned information about Go is not explicitly part of its design, nor is its information accessible from the perspective of someone using the design stance. I can potentially

[99] Note that I do not write 'win Go' as the goal or design function. Often, neural networks being trained has to learn the main goals or victory condition themselves through playing the game over and over again.

say that AlphaGo made X and Y moves, because they were the objectively 'best' moves at the time or because it was a winning strategy, but no explanation deeper than that is available. As such, the design stance, while it is usable for both prediction and explanation of AlphaGo – and certainly an improvement on the physical stance – its answers lack the finesse and nuance used for a more detailed analysis of behaviour.

If the design stance is all I have to go on when trying to predict my opponent's moves – then, I will make a rather poor show of it. Interestingly, the same would be the case for any of AlphaGo's makers playing against it. Even intimate knowledge of AlphaGo's design will not yield much success when playing against it. In other words – knowing AlphaGo's design at any level gives no advantage in attempting to predict its moves. Simply put, what AlphaGo was designed to do was not to win at Go by pursuing such-and-such strategies, but rather to learn 'on the job', somehow or other, *how* to win. So even knowing plenty about AlphaGo's actual design will not help me much in predicting or explaining its specific moves.

Thus the physical stance helps almost naught, when trying to explain and predict my opponent's outputs, and the design stance, while better, only gets us so far. I shall make a further claim here though – namely that the style of explanation and prediction that the design and physical stance can provide simply does not match the stance that is actually used by people facing autonomous systems. Imagine me again in the middle of my Go match against AlphaGo. Facing the puzzle of trying to predict my opponent's next move, how would I actually go about it?

I claim that – knowing as I do what a fearsome opponent AlphaGo is – I would go about it in the same way as I would approach the issue if faced with an excellent human opponent. To predict my opponent's moves, I will ask questions such as: 'Which strategy/series of moves would make the most sense for it to execute considering the state of play?', 'What is it up to – is it

trying to trick me? Is it planning and aiming to trap me?', 'Does it see through my plan?' or 'Since I've made this particular move, it might think I want to pursue a certain strategy, so how would it block it?'. These kinds of questions clearly belong to the intentional stance: I am attributing intentional states to the system and viewing it as a rational Go-player.

I can therefore get a lot further with the intentional stance than I can with the design stance. I am able to sensibly think something like: 'Aha! With that move, AlphaGo is trying to trick me into weakening the left side of my board!'. Or I might say: 'AlphaGo places its game piece at X instead of Y because it knows I will otherwise put my piece at Z and AlphaGo is afraid of being locked in', or 'AlphaGo put its piece there because it wants to finish the endgame fast'. In other words, the intentional stance allows me to explain and predict AlphaGo's moves with much more nuance; and this, in turn, will enable me to play better against it.

The intentional stance, then, allows me to see AlphaGo as a strategic opponent. Using the intentional stance, I can say that AlphaGo is strategizing, trying to trick and outmanoeuvre me. This is the punch that the design stance was missing. Only by seeing the game in this way am I able to counter its cunning plays and I stand a better chance against my opponent. Hence, in order to read and predict my opponent's moves the best, I have to play my game of Go against AlphaGo in the same manner as I would against a human being – ascriptions of cunning intentions, strategic tricks and all. This is the intentional stance in action. In the heat of the battle, I do not waste time trying to figure out impossible calculations or considering design features, when attempting to predict my opponent's moves. Instead, I go straight for the mind games: I ascribe rationality, reasons, and intentional states to my opponent, as I try to claw my way to an unlikely victory.

The above example shows how one might use the intentional stance in a situation featuring an autonomous system. I here used Google DeepMinds' AlphaGo to explain the

intentional stance and how it works in practice. I did this because most people will have encountered systems similar to AlphaGo, albeit those will have been systems of lesser complexity. Some brief examples might be online chess AIs or even food-delivering drones, which I will talk about further in the following chapter. Of course, the intentional stance is not only apt when used to explain or predict the moves of game-oriented systems such as AlphaGo. The above argument may be extended to capture situations, where humans seek to explain or predict the behaviour of any learning autonomous system. More examples of this will follow in Chapter 6.

The AlphaGo example has illustrated how the intentional stance works with an autonomous system. More than that, the example has shown how the intentional stance is not just a possible explanation of machine behaviour, but indeed the best and most useful explanation compared to the other stances – as is shown by the fact that, plausibly, it is the stance we actually use when faced with predicting and explaining the behaviour of such systems like AlphaGo.

This is then the lay of the land. The intentional state is the best tool for explaining machine behaviour through the ascription of intentional states: states that collectively rationalise behaviour. Furthermore, since the intentional stance is applicable to autonomous systems (as it was shown in the Go example that it can be), the ascription of rationality – or acting 'for reasons' – is legitimate. Therefore, this provides the answer to the 'inside-the-box' question posed earlier, namely: how can we attribute understanding of reasons to something that is seemingly nothing more than data, sensors, wires and flashing lights? On the Dennettian approach I have adopted here, 'understanding of reasons' simply amounts to the appropriateness of the intentional stance in predicting and explaining a system's behaviour.

I stress that on this broadly Dennettian view there is no requirement on the system that it be conscious or intelligent in a human sense or anything else – its physical constitution, over

and above how that physical constitution delivers behaviour that is best explained and predicted using the intentional stance, is irrelevant. The ascription of intentional states and reasons to autonomous systems just hinge on the intentional stance being the best applicable stance for explaining their outputs. So, my response to worries about the 'inside-the-box' matters simply boils down to the legitimate use of the intentional stance when dealing with autonomous systems. And according to the intentional stance, the story – and the possible accompanying questions or objections about the 'mental' or inner aspect of autonomous systems – ends there.

That being said, the sceptic might still raise some immediate objections to the use of the intentional stance for autonomous systems. I will here consider two possible objections. The first builds on scepticism about the general practise of using the intentional stance with non-human things in general. The objection might run along these lines: It is all very well and good that the intentional stance is the best explanation for the behaviour of autonomous systems, but humans anthropomorphise objects of all types all the time – how is this different?[100] One may easily imagine that I can stick some googly eyes on my pencil and name it Patrick, and I might then feel disinclined to sharpen him, as I do not want to hurt him. Going down that route, I can easily attribute a whole host of beliefs, thoughts and feelings to Patrick. How do we know that when using the intentional stance with autonomous systems, we are not just inappropriately anthropomorphising them? How is it any more appropriate or apt than it is with Patrick the pencil? More specifically, if using the intentional stance with autonomous systems are mere cases of ignorant anthropomorphising, then the intentional stance is not necessarily the best stance for dealing with such systems, thereby rendering it inappropriate as an answer for the 'inside-the-box' matters relating to autonomous systems and reasons-responsiveness.

[100] See as an example Kim and Shyam Sundar 2011 for an experiment showing that human agents tend to unconsciously anthropomorphise non-autonomous machines.

I will argue that this objection fails, as there are two main things that differentiate the use of the intentional stance in relation to autonomous systems from the general anthropomorphising of objects. First, the intentional stance (on the interpretation used here) is an aid to explain and predict the behaviour of objects and agents. The very purpose of the stance thus limits the types of objects that can be sensibly viewed using this stance: it can only be sensibly used with objects or agents that display some kind of behaviour, usually in the form of action or outputs. The first distinction between anthropomorphising objects and using the intentional stance therefore lies in the difference in the range of objects that these concepts can be applied to.

Patrick the pencil does not exhibit any behaviour – and even, if one for some reason would argue otherwise, it would not be 'behaviour' that is best explained by appeal to intentional states. When Patrick deposits lead on the page in a way that exactly matches what I want to write, the explanation according to which he is dutifully obeying my wishes, or is sensitive to my desire to write a shopping list, or figures that it is in his best interests to do what I want him to do, is no better than the explanation according to which he is merely transferring lead onto the page thanks to the pressure I am exerting on him according to straightforward physical laws.

The second distinction lies in the differentiation in purpose between anthropomorphism and using the intentional stance. The signature characteristic of everyday anthropomorphism is personification of objects rather than the attribution of individual mental features to them. Whether the given objects are completely lifeless or do perform some simple given tasks (say, like those of a Roomba), the core purpose of anthropomorphising is constructing a narrative, where the object has personhood or human-like subjective experiences of some kind. Hence, anthropomorphism can be done to almost any type of object – it is not about trying to make sense of the objects output or explain or predict its behaviour, but instead to tell a narrative of some

kind. Suppose as an example that I put googly eyes on my Roomba and call him Fido.[101] Fido likes to get petted, when he comes over to get his sensors cleaned, and I can say that Fido has grown lazy in his older years. I might say that Fido likes watching Midsomer Murders with me, though he has never quite warmed up to the new Barnaby. All these things help establish Fido as a 'character' or a thing with a loosely defined personality. However, none of these anthropomorphising statements are part of an explanation of the Roomba's outputs, nor are they useful in predicting what the robotic hoover is going to do next.

The key point that I wish to make is therefore that in anthropomorphising an object, whether that be a Roomba, a pencil or something completely different, one engages in a kind of generalised pretence whereby we treat it as a person and explain *all* of its behaviour in intentional terms, whether or not doing so constitutes the *best* explanation – which, of course, it normally doesn't. Perhaps if I describe Fido as having gotten lazy in his old age, I might intend that to be an explanation of why he cleans less efficiently and effectively than he used to. However, a better explanation, not least because it would lend itself to improve Fido's future efficiency, is just that his dust filter or motor needs replacing. In contrast, adopting the intentional stance is something that is done, or can be done, on a case-by-case basis depending on the specific behaviour that needs to be explained or predicted: different stances might be appropriate with respect to only some behaviour on some occasions. As an example, if AlphaGo stops responding due to lack of connection or due to a loss of power, then explaining its unresponsiveness by appealing to the system being bored or tired is not helpful. If and when it is not useful, the intentional stance may be ditched in favour of a better stance – design or physical – for explaining the given situation. Anthropomorphising simply isn't aimed at explanation or prediction at all.

[101] The example is inspired by the common practise amongst Roomba owners of treating their robotic hoover as a dog.

The second potential objection concerns the use of Dennett's intentional stance in order to answer the 'inside-the-box' worry. As was mentioned earlier in this section, assuming Dennett's intentional theory is a controversial move, and as such one might object that this move does not really answer the inside-the-box question. Hence, a sceptic might worry that the machine compatibilist account is resting on shaky ground.

I will argue that any such worry is unfounded given my purposes. The machine compatibilist account does need a coherent story of how reasons for outputs might be attributed to autonomous systems, yet there is no requirement on what that story should look like and as such an appeal to the intentional stance is just one way that story might go. One could in principle exchange the current 'inside-the-box' story for another without it having a knockdown effect on the rest of the account.

An obvious contender for a substitute theory for the 'inside-the-box' matters would be something from the functionalist camp. Functionalism is a theory of mind according to which a mental state is not defined by its internal constitution, but instead by its function. According to a functionalist account, what exactly makes up the 'mental' aspect of an autonomous system is irrelevant, as long as it serves the right functional uses. More specifically from the functionalist camp, 'machine functionalism' would probably be an obvious choice for the purposes of this thesis.[102] The point being that there are other options for accounting for the 'mental' aspect of machines. So even if one is suspicious of Dennett's view, one need not worry about the integrity of the machine compatibilist account as a whole – that particular aspect of it is exchangeable.

[102] See Levin 2021 and Putnam 1967 for an introduction to functionalism and machine functionalism. See Block 1978 for some of the most pressing obstacles to such an account.

Using Dennett's theory does come with its own benefits though. Notably, one of these is, of course, the lack of need for 'internal' analysis. When working with autonomous systems, where each new generation of system changes and differs inside, there is an obvious benefit to using a theory that does not require the identification of specific internal aspects of the system and their functions. However, if one insists on rejecting the Dennettian view, the machine compatibilist account has been shown to basically just need an alternative 'machine-positive' story of internal matters. As this thesis is trying, overall, to develop an account for morally responsible autonomous systems in response to the moral responsibility gap, this seems like a fair price.

So then, it all comes together. How might we attribute reasons to and explain the behaviour of autonomous systems? It turns out that the intentional stance is applicable to autonomous systems, specifically it turns out that the intentional stance gives the best explanation of certain autonomous systems' behaviour. With the intentional stance, intentional states and in turns rational goals, reasons and the likes are attributed to autonomous systems. Going forward, when referring to reasons in respect to autonomous systems, I use the term in reference to the sense that it has been presented in this section.

Chapter 6: Limits and Manipulation

In Chapter 4, I showed that guidance control consisted of two aspects: moderate reasons-responsiveness and some form of a no-manipulation condition. In Chapter 5, I then argued that even simple autonomous systems can generally be ascribed moderate reasons-responsiveness. In this last chapter, I will discuss the no-manipulation condition for guidance control and detail how this condition can take form, thereby completing the outline for my reasons-responsive machine compatibilist account. I will argue that despite the fact that autonomous systems can be ascribed moderate reasons-responsiveness, both current and near-future systems fail to fulfil the no-manipulation condition of guidance control, thereby ultimately failing to fulfil the freedom-relevant condition for moral responsibility. I will then argue that, nevertheless, my reasons-reactive machine compatibilist account can be used to answer the conundrum of moral responsibility gaps in contemporary and near-future case scenarios. To achieve all this, the chapter will be divided into five sections.

In §I, I will argue that autonomous systems routinely have their reasons-receptivity purposefully limited. In §II, I will then show that this limitation on reasons-receptivity is categorised as manipulation when found in human agents. I will then argue that this specific type of limitation of reasons-receptivity, when applied to human agents as well as autonomous systems, exempts them from moral responsibility for their actions/outputs. In other words, I will argue that autonomous systems qualify as manipulated entities, excusing them from full attributions of guidance control and in extension, moral responsibility.

In §III, I will then argue that in the above-mentioned manipulation cases, the manipulator can be found to be morally responsible for the given action or output. I will further

argue that even if the given action or output is unwanted or unintended by the manipulator, it does not relieve them of the moral responsibility for that action or output.

In §IV, I will discuss the possibility that creators and/or users are morally responsible for the outputs of autonomous systems. I will show that each of these parties play a specific role in the 'manipulation' process of autonomous systems and I will argue that in light of these roles, the users can be found to be morally responsible for the outputs of the autonomous systems they use and for the immediate harm that may come about as a result of these outputs.

In §V, I present a few classic examples of moral responsibility gap problems, as introduced in Chapter 1, and show how these are solved by using my machine compatibilist account in conjunction with the previous sections' findings on manipulation. As such, while I in §IV argue that the user is morally responsible, I in this section show how this conclusion is applied in practise, when encountering general moral responsibility gap problems. Hence, this section will show how this thesis' account can be used in practice to solve the conundrum of moral responsibility gaps for current and near-future scenarios, thereby solving the puzzle of moral responsibility gaps as it was set out in Chapter 1.

I. Limitations on Reasons-Receptivity

Back in Chapter 4, I established that fulfilling a no-manipulation clause was needed together with moderate reasons-responsiveness in order to achieve guidance control. However, it was also made clear that F&R's original attempt at such a clause, in the form of a condition on mechanism ownership, failed to negate newer and more sophisticated manipulation worries. I will in this chapter return to this topic and discuss the worry of manipulation in relation to moderate reasons-

responsive autonomous systems. In these two first sections I will argue that despite being reasons-responsive, most autonomous systems fail to fulfil the freedom-relevant criteria for moral responsibility due to the presence of a particular type of manipulation. To start, in this section I will argue that there is an oddity to be observed when taking a closer look at the reasons-receptivity of autonomous systems.

I showed in the previous chapter that autonomous systems can easily be moderately reasons-responsive in line with F&R's definitions: I argued that even a simple autonomous system such as a turtle is both strongly reasons-reactive and regularly reasons-receptive, hence easily fulfilling the conditions for moderate reasons-responsiveness. Nevertheless, in §III of Chapter 5, I discussed briefly a worry about the limited range of reasons that an autonomous system could be receptive to; specifically, I talked about their obvious lack of receptivity to moral reasons. In Chapter 5, I argued that the objection fails because the lack of receptivity to moral reasons in particular does not preclude the attribution of reasons-receptivity to autonomous systems, however I will here expand on this potential worry as I will show that a lack of receptivity to a wide range of reasons can nevertheless affect attribution of moral responsibility.

While regular reasons-responsiveness can be attributed to autonomous systems, I will argue that contemporary and near-future systems are only receptive to a very limited range of reasons. To illustrate, consider an autonomous cleaning system. One might imagine that the robot has been given an overarching goal, namely cleaning a specific room whenever it is needed. Being a sophisticated autonomous system, suppose that the system sets its own smaller specific tasks to fulfil its goal. As such, the system gets input from observing the state of the room and chooses a cleaning agenda based on this input.

Suppose that the system registers the floor to be dusty and that this input is sufficient for the system to initiate vacuuming. The question is then: is the system reasons-responsive in this scenario? From what has been presented in the previous chapter, the answer is yes. There are nearby possible worlds where the dust levels are different, and where the system registers this input, resulting in different outputs. For example, there is a possible world where the windows are dirtier than the floor and so the system registers this as a sufficient reason to do otherwise and hence tackles the windows instead of the floor. For the cleaning robot, all the possible inputs that the system is receptive to correspond to nearby possible worlds, where the system (being strongly reasons-reactive) reacts to these. Insofar as the cleaning robot can be attributed strong reasons-responsiveness and regular reasons-receptivity, the system is moderately reasons-responsive.

Contrast now this with a similar scenario, but instead of a cleaning robot being given the task, you ask a teenager to clean the same room. Suppose that the young adult in this scenario also observes the dusty floor and recognises the state of the floor as a sufficient reason to vacuum, which he then proceeds to do. As such, in the two scenarios both the autonomous system and the teenager have the same overarching task, i.e. cleaning the room regularly, and they both react the same way upon recognising the floor as sufficiently dusty.

The teenager in this described scenario may also be found to be moderately reasons-responsive. It may be assumed that the young agent exhibits a sensible pattern of recognising reasons to do otherwise in a range of nearby possible worlds. Similarly, it may also be imagined that the agent does act upon one of these reasons in at least one of the nearby possible worlds. Depending on these conditions, the agent can be considered moderately reasons-responsive with respect to vacuuming the floor.

On paper, we have two reasons-responsive agents/entities in scenarios that are seemingly indistinguishable with respect to their reasons-responsiveness. Both the cleaning robot and the teenager perform the same action/give the same output across a range of nearby possible worlds and are therefore both reasons-responsive in relation to the action/output. Yet, these two cases will strike most readers as unequal. I will argue that this intuition is right and can be explained by a difference in the scope of reasons-receptivity.

Both the autonomous system and the teenager are moderately reasons-responsive. But, consider now the scope of reasons that the teenager is receptive to in nearby possible worlds. Without changing anything about the teenager as a person or as an agent, he may be imagined to be receptive to a whole host of reasons. In one nearby possible world, he might see the windows and register them as in need of a wipe-down. Just like the autonomous system in the alternative sequence of events described above, we can assume that that the agent sees the state of the windows as sufficient reason to clean them immediately.

In another possible world, however, he might notice an old photo album next to the vacuum cleaner, and see the opportunity to reminisce over old times to be a sufficient reason to sit down and flip through the pages. Or, in another nearby possible world, he might see that his sibling is upset and considers this to be a sufficient reason to go comforting them instead of vacuuming. The number and range of possible reasons that a normal human teenager is usually receptive to is vast.

When considering the scope of reasons-receptivity for the autonomous system described earlier, the same cannot be said. Without significantly altering the history of the autonomous system, there will be no nearby worlds where the system gets distracted by memories or chooses to venture away from the task because someone is upset. By design, the cleaning robot is

only receptive to a set of reasons that specifically relates to its given tasks. As such the set of possible reasons the system is receptive to is incredibly limited compared to the human agent.

The scope or range of the system's reasons-receptivity is limited by its sensors and initial programming. Consider the autonomous turtles, as described in Chapter 5. A turtle is equipped with a light-sensitive sensor and nothing more. It will not react to loud noises or colourful displays – only light. Though much more sophisticated, the cleaning robot in this scenario is fundamentally similar to its turtle ancestor. In the described scenario, the system might be imagined to be receptive to a comprehensible list of reasons for doing otherwise. As mentioned, it might register the windows to be dirty, or the rug in need of a wash or similar. However, all the possible reasons that the system might be receptive to will be directly linked to its cleaning task, the system's own maintenance or movement. As such, the range of possible worlds where the system reacts to reasons for doing otherwise, is limited by its receptivity to this narrow list of inputs.

One should not infer from this limitation of receptivity that autonomous systems are not regularly reasons-receptive in the way needed for moderate reasons-responsiveness after all. Recall from chapter 4, the example of Charis, who buys herself an ice-cream cone. She fulfilled the conditions for moderate reasons-responsiveness despite only recognising one dimension of reasons for doing otherwise in nearby possible worlds. As such, she was deemed regularly reasons-receptive due to her sensible pattern of reasons-recognition in the possible worlds, where the queue to the ice-cream van was entirely too long. Hence, though Charis' reasons-recognition was one-dimensional, insofar as she was just receptive to the different queue lengths as reasons to not get an ice-cream cone, she nevertheless fulfilled the conditions for moderate reasons-responsiveness. As such, the limitation of the range of reasons that an agent can be receptive to does not undermine an agent's

being regular reasons-receptive, as long as the agent still exhibits a sensible pattern of reasons-recognition with respect to the reasons that they are in fact receptive to.

Now, let us return to our cleaning robot and the limited range of reasons that it is receptive to. One might question whether the system really has such a limitation by pointing out that there are other possible worlds where the system has different sensors, and so the system is in fact receptive to the broader range of reasons corresponding to the different inputs those sensors would enable. However, this is much like claiming that, since there are other possible worlds where a colour-blind human agent has a visual system that allows them to discriminate between red and green, the agent is actually receptive to reasons that relate to redness vs. greenness. Both the system and the human agent are indeed receptive to additional reasons to do otherwise in such possible worlds, where their key features have been changed. Nevertheless, recall that when ascertaining the reasons-responsiveness of a mechanism by looking at nearby possible worlds, the mechanism in question must be held fixed. By considering possible worlds where sensors have been added, one fails to hold the mechanism fixed and therefore these worlds are irrelevant when considering the reasons-receptivity of an agent or system.

In this section, I have argued that there is an intuitive difference between a human agent and a general autonomous system, even when the two are placed in identical sequences of events and are both moderate reasons-responsive with regard to their action/output. I have further argued that the difference between the two can be explained by a difference in scope or range of reasons-receptivity. In the following sections, I will discuss further what consequences a limited scope for reasons-receptivity has for the attribution of moral responsibility to an agent.

II. Autonomous Systems as Manipulated Agents

In this section, I will argue that the severe limitation of an agent's reasons-receptivity affects attribution of moral responsibility and may amount to a type of manipulation. To do so, I will present a distinction by Yaffe (2003) between coercion and indoctrination as manipulation methods. I will then argue that intentional limitations of the scope for an autonomous system's reasons-receptivity is a form of manipulation akin to indoctrination. If this type of manipulation intuitively keeps human agents from being considered morally responsible for a given action then, I will argue, the same must be said for manipulated autonomous systems.

Yaffe (2003) distinguishes between two different types of manipulation: coercion and indoctrination. The difference between the two can be illustrated with a few examples. Imagine first a case of somebody being blackmailed. The agent is told that he has to steal a valuable jewel or else harm will come to his family. Our agent in question is well aware of the many reasons not to steal the jewel, yet he still does it at the blackmailer's behest. Yaffe writes about people in this kind of situation: 'Such a person feels the force of the other's manipulation as doing violence to her agency; the manipulator causes her to lose a battle to the motives on which the manipulator wants her to act' (2003, 338). This constitutes the first type of manipulation: coercion. Coercion covers most of the traditional types of manipulation, the gun-to-the-head kind of scenario. There is typically a clear force, which the manipulated agent is aware of and to which they succumb if the manipulation is successful.

The second type of manipulation is indoctrination. Imagine a case where an agent has grown up in a cult with a tremendously charismatic leader. Suppose that one year the cult leader successfully manages to convince his followers that world is going to end. He says that the only

way for his followers to find salvation is to give all their material possessions and riches to him. Our agent does the cult leader's bidding.

A successfully indoctrinated agent shows a particular limited pattern of responsiveness to reasons as a result of the manipulation. The pattern or range of reasons that the indoctrinated agent is responsive to is usually that wished by the manipulator. Yaffe writes about the indoctrinating manipulator: 'By manipulating them, he causes them to recognise and respond to those features of the world which give *him* reason to have them act in particular ways' (2003, 340).

The indoctrinated differ from the coerced, as they have no experience of being forced or restricted by their manipulators. No threats or force are used to get the indoctrinated to behave according to the manipulators' wishes. As such, it might at a first glance be difficult to say exactly how the manipulator restricts the indoctrinated agent. After all, one might be tempted to look at the cult followers and think that it is a shame, but they are not being forced to follow the whims of the cult leader. Nevertheless, it is intuitively true that an agent can be severely indoctrinated to such an extent that one may deny that they are morally responsible for their manipulated actions. For the remainder of this thesis, I will solely focus on indoctrination-style manipulation.

Within the compatibilist literature, there is a large debate trying to differentiate large scale manipulation, such as indoctrination, from the standard constraints of being an agent in a determined system.[103] Yaffe (2003) argues that the difference between the indoctrinated agent and the mere determined one is found when comparing the nearby possible worlds for the two of them. He writes: 'When we fall into the hands of indoctrinators, fewer lives are available to us than are available to us when we are simply the unlucky victims of neutral causal forces' (Yaffe 2003, 345).

[103] See Mele 2006 and Pereboom 2001 for examples.

I will here explain and discuss this idea in further detail using F&R's concept of reasons-responsiveness and its components.

Let us start by recalling the cult example. Imagine two agents who both give over their material wealth to the cult leader. Suppose the first agent's actions are a result of long-term manipulation, while the second agent's actions are merely a result of an unfortunate pre-determined fate. As such, one agent is indoctrinated and the other just a standard determined agent. I will here show how these two agents differ from each other. Now, let us first suppose that we wish to find out if these agents are moderately reasons-responsive in respect to their action of giving all they own to the cult leader. From F&R's account, presented in Chapter 4, we know to look towards nearby possible worlds for an answer. Recall that to have guidance control according to F&R moderate reasons-responsiveness is needed. As such, what is needed is demonstrable weak reactivity to reasons and regular reason receptivity. Suppose that both agents turn out as moderate reasons-responsive in respect to their action (in line with such examples as Charis and her ice-cream cone purchase in Chapter 4), and as such any further questions about their reasons-responsiveness can be parked – it is not on this aspect that the agents differ from each other. While their moderate reasons-responsiveness is not in question, Yaffe argues that when considering nearby possible worlds, it will still be the case that there are certain counterfactuals that will be false for the manipulated agent, while true of the determined, hence a difference between the two can be drawn.

To view the difference, let us consider some nearby possible worlds for the determined agent and the manipulated agent respectively. To start, consider a range of nearby possible worlds, where the determined agent leaves the cult compound and talks to family members or old friends. Recall, that the determined agent is in the cult purely due to being dealt a poor hand by his determined fate, and as such the cult leader has no special interest in this agent. As the

determined agent leaves the compound and listens to people outside of the cult, we may presume that the determined agent in some of these worlds recognises reasons to not give all his things to the cult leader and may even act upon these reasons in some of these worlds. In other words, if other possible worlds presented our average moderately reasons-responsive determined agent with a new range of reasons to do otherwise, then all else being equal, his mechanism for decision-making would be receptive to the presented reasons. As such, being a determined agent alone does nothing to limit one's receptivity to different ranges of reasons in nearby possible worlds.

However, consider now relevantly similar nearby possible worlds for the manipulated agent. Here, the cult leader plays as important role as the manipulator. If a family member or friend were to try to get in contact with the manipulated agent, the manipulator would simply seek to prevent this. As such, the cult leader might keep the agent from leaving the compound, take away their cell phone, or something similar. In other words, in a wide range of nearby possible worlds, the manipulator interferes in such a way as to avoid the manipulated agent acting differently than from the actual world. More specifically, the manipulator in such nearby possible worlds seeks to limit the agent's access to reasons to do otherwise, thereby limiting the manipulated agent's reasons-receptivity to the reasons that align with the manipulator's wants. Herein lies the difference between the standard determined agent and the manipulated agent. In nearby possible worlds, the manipulator interferes with the manipulated agent's access to reasons. In contrast, as Yaffe puts it: 'The Unlucky [read here: the determined agent]… would simply stray from the course and come to have a different pattern of response to reasons from that of the Manipulated' (2003, 343-344).

Of course, there are nearby possible worlds, where our manipulated agent does otherwise – where the safe holding the agent's phone is unlocked or where the compound guards are having a nap on the job and hence gets access to family or friends who presents them with

reasons to do otherwise. However, these possible worlds are all ones in which the agent – by luck – somehow breaks the pattern of reasons-responsiveness instilled by the manipulator. In other words, these possible worlds are those where the attempted manipulation is unsuccessful: it is somehow not affecting the agent.

The point here is this. On the one hand, our manipulated agent is regularly reasons-receptive, because – for a very limited range of reasons to do otherwise, namely those that the manipulator might have for *wanting* the agent to act otherwise, in nearby possible worlds where there are such reasons, the agent will be aware of them and will, as a result, do otherwise. For example, were the manipulator to ask the agent to do something else today – guard the compound, say, or just transfer half of all their money to the manipulator so as not to attract too much attention to what's going on – the agent would receive those reasons to do those things, and they would do them.

On the other hand, the *range* of reasons that the agent is receptive to – the boundary, as it were, of their reasons-receptivity – is entirely set by the manipulator. Worlds in which the agent comes to realise, for example, that in fact the world is not going to end and so decides not to give the manipulator all his money, are not the worlds that are accessible to him: the manipulator has seen to that. To put it in Yaffe's terms, there are some 'available lives' for the non-manipulated agent that are not available lives for the manipulated one (2003, 345). My claim, then, is that the manipulation consists not in rendering the manipulated agent fail to be regularly receptive to reasons, but in rendering the *range* of reasons to which the agent is receptive too narrow for them to be morally responsible for their actions. They have moderate reasons-responsiveness, but of a kind that is too limited for it to be appropriate to ascribe moral responsibility to them.

Recall now the scenario from the previous section featuring a cleaning autonomous system and a young adult both vacuuming a room. I will argue that the difference between them can be explained by the autonomous system qualifying as a type of manipulated agent akin to indoctrination, whereas the teenager in the case is under no manipulation at all.

Recall that we found the young adult to exhibit receptivity to wide range of reasons. Whether the reasons be that the agent has urgent homework to do, a sudden tornado is coming towards the house or even just the agent registering the sequence of events as an ideal opportunity to rebel against their caregiver, in nearby worlds where these reasons obtain, our teenager will be sensitive to them and, we may assume, act accordingly. As long as the agent has a perfectly ordinary decision-making mechanism, he is regularly receptive to reasons. Furthermore, crucially, he is also receptive to a very broad range of potential reasons to do otherwise. Of course, he is not receptive to *all* potential reasons to do otherwise. For example, in a nearby possible world where the hoover looks like it's functioning normally but in fact has an electrical fault that's going to make it burst into flames when it's switched on, there is an excellent reason to do otherwise that the teenager isn't aware of. So, the teenager (in the actual world) isn't receptive to *that* possible reason to do otherwise. But he is receptive to a very large range of such reasons.

Compare then the teenager to the cleaning robot. As I explained in the previous section, there is a difference between the standard agent and the cleaning robot: the cleaning robot is receptive to a much narrower range of reasons. As such, I suggest that the cleaning robot is different from the standard determined agent because it has, in Yaffe's (2003) words, 'fewer available lives'. As in the cult case, the cleaning robot differs from the teenager in virtue of having severe restrictions on the range of reasons it is receptive to.

The cleaning system is receptive only to reasons directly linked to its cleaning tasks and for which it has corresponding sensory mechanisms. As such, it is reasons-receptive to such things as dirty windows, dusty bookcases, and full trashcans. Within its limited area of operation, the system is moderately reasons-responsive. Yet, consider a nearby possible world where a person next to the robot is having a stroke, or there is a tornado on its way. In the respective possible worlds where these events occur, there are sufficient reasons for the robot to do otherwise than clean, such as alerting someone to person having a stroke, getting out of the tornado's way and so on. Yet, without changing the build and sensors of the cleaning robot drastically, the system is not receptive to these possible reasons for doing otherwise: in the possible worlds where these things happen, the robot is completely oblivious to them. The examples of the person having a stroke and the tornado are fairly dramatic examples, of course, but any reason outside of the system's task area – and outside the realm of the sensory equipment it has been built with, for the purpose of carrying out those tasks – will serve just as well as an example. A non-manipulated agent will, as a general rule, usually not be especially limited in the range of potential reasons to which they are receptive. No actual agents are completely unlimited in this regard, of course, but as was also seen when considering the cleaning teenager earlier, normal humans are receptive to a vast range of potential reasons. The contrast to the cleaning robot here should therefore be clear. Not only is the cleaning robot's range of reasons-receptivity limited, but it is severely so. As the autonomous system is severely limited in the range of reasons it is receptive to, it mirrors the indoctrinated agent. As such, the cleaning robot may be viewed as a form of manipulated agent.

Recall from Chapter 4, §IV that F&R's compatibilist theory as well as my machine compatibilist account requires two conditions to be fulfilled in order for an agent/entity to be ascribed guidance control and thereby fulfil the freedom-relevant condition for moral responsibility. The two conditions are, of course, moderate reasons-responsiveness as well as some 'no-

manipulation clause'. The latter condition is meant to prohibit ascription of guidance control to manipulated agents, as manipulated agents intuitively are not 'free' in the relevant sense for moral responsibility. I have not here attempted to give an analysis of what this 'no-manipulation clause' consists of. However, I have argued that indoctrinated agents will not fulfil such a clause, and I have shown that contemporary and near-future autonomous systems are limited, when it comes to the range of reasons they are receptive to, in the same way as an indoctrinated human agent, and claimed that this constitutes a form of manipulation – though doubtless not the only form. Given this, current and near-future autonomous systems will clearly fail to fulfil the 'no-manipulation clause' and thereby fail to fulfil the conditions for guidance control, despite being moderately reasons-responsive.

Recall the autonomy levels from Chapter 5, §I. There, it was shown that current and near-future autonomous systems are sometimes categorised as having different levels of autonomy. Level 5, being the highest, requires the autonomous system to function fully without any need of human supervision. Imagine a level 4 or 5 surgical robot. Such a system would be highly sophisticated. One might imagine it to be able to diagnose patients alone, decide that the patient needs a specific type of surgery and perform the surgery, all without human oversight. However, even an autonomous system at the highest level of autonomy would still be severely limited in the range of reasons it is receptive to in comparison to the standard human agent. It then must be clear that for current autonomous systems, not only a high level of autonomy but a much wider range of reasons-receptivity is needed before one can have further discussions about the possibility of these systems fulfilling the freedom-relevant conditions for moral responsibility.

Failing the 'no-manipulation clause' means that autonomous systems cannot fulfil the freedom-relevant conditions for moral responsibility. But nor are they mere tools: we are dealing

with moderately reasons-responsive entities, albeit manipulated ones. If autonomous systems can be put in the category of indoctrinated/manipulated agents, it opens up questions about the identity of their manipulators and, relatedly, questions about who is morally responsible for the consequences of these manipulated entities' outputs. This is the topic of the following section.

III. The Moral Responsibility of Manipulators

In this section, I will argue that in cases of indoctrination for both human agents and autonomous systems, the manipulator is morally responsible for the indoctrinated agent's actions/outputs. I will further argue that this conclusion also extends to cases, where the manipulator does not have full foresight of what the indoctrinated agent will do. Before we begin, I would like to apologise in advance for the use of rather macabre themes for this section. The examples have been made so in order to present actions and consequences that carry intuitive and clear moral value.

Let us start by considering cases where the indoctrinated agent acts according to the wishes of their manipulator. In classic philosophical literature, the manipulator is easily identified. There, the manipulator often takes the form of an evil doctor, a cruel neuroscientist or a deceitful cult leader. In §II, I introduced a case featuring a cult leader, so let us continue with that here. Consider again the indoctrinated agent who only recognises reasons for action when they are reasons for actions that benefit the cult leader. In other words, through clever indoctrination the cult leader has made sure that the indoctrinated agent does not have access to a range of reasons for doing otherwise, and therefore the agent cannot recognise these reasons in the actual sequence of events.

Recall from the cult example in §II that the indoctrinated agent hands over all of his belongings to the cult leader. Let us now add a bit to this story. Suppose that the cult leader has his indoctrinated followers fully convinced that there is a better life after this earthly one – such that this life is merely a steppingstone to eternal bliss. The cult leader himself believes there to be no such thing as an afterlife; instead, he believes only in the power and desirability of material wealth. The cult leader has made sure that his indoctrinated followers are sensitive to the reasons for actions which benefit him. Further, though the cult leader would interfere such as to secure the cult followers' limited receptivity to a range of reasons in a series of nearby worlds (just as was discussed in §II), we may imagine that he needed not do any such thing in this actual sequence of events. Nobody leaves the compound, gets calls from relatives or similar things that might give the cult followers access to a range of reasons to do otherwise. As such, the cult leader's plan goes off without a hitch, as the indoctrinated agents end their own lives while leaving him with all their worldly goods. This example might be grisly and extreme, but unfortunately not unrealistic, as the world has been witness to cult tragedies such as this before, where indoctrinated people end their own lives at a cult leader's behest.[104]

Who is morally responsible for the followers' suicides and their leaving their material things to the cult leader? I will argue that that this responsibility falls on the cult leader. Take our indoctrinated agent again – his categorisation of being indoctrinated is based on his limited access to reasons to do otherwise in nearby possible worlds and therefore the limited range of reasons to which he is receptive. This limitation in question is done by the cult leader.

[104] While the Jonestown massacre might be the most famous case of cult suicides, the mass-death incidents featuring the Order of the Solar Temple is more like the case described here.

If the cult followers are categorised as indoctrinated agents, then this categorisation is based on their reasons-receptivity having been limited. The limitation in question is done by the cult leader through indoctrination. By indoctrinating another agent, the manipulator alters or manipulates the range of reasons that the indoctrinated agent is receptive to in order to further their own agenda. While the indoctrinated agent may be moderately reasons-responsive, the constraints or limits on their access to reasons is a testament to the manipulator's use of them as a tool for their own purposes. As such, from the point of the view of the manipulator, the indoctrination is done to ensure that their will or agenda comes through.

In the case featuring the cult leader, the indoctrination was done with the purpose of robbing his followers of their material wealth. He limited his followers' possible access to reasons to do otherwise in order to ensure a scenario where the deaths of his followers and the transfer of material wealth would happen. It seems intuitively fair, then, to say that the cult leader is morally responsible for the events that his indoctrination brought about.

So, it has been argued that manipulators can be considered morally responsible for the intended events caused by their indoctrination of their followers. Now, I will go a step further and argue that manipulators may also be considered morally responsible for actions that happen due to their indoctrination, but that are not necessarily intended nor even wanted. To do so, I will specifically argue that agents are responsible for the outcomes of their moral gambles, including ones that involve indoctrinated agents.

To show this, I will start off with a much simpler example. Imagine that an agent, Mary, has a friend who is in terrible financial trouble and will be in serious trouble if their debt is not paid off by tomorrow. Mary wants to help and has an inside tip that a certain horse will be a sure winner at the local races. Suppose that Mary secretly gets access to her friend's savings and

bets it all on the aforementioned horse. She is not being pressured to make this bet, nor is she herself being manipulated in any relevant manner. We may assume for the purpose of this case that she is a normal, moderately reasons-responsive human agent.

Mary bets the money on the horse with the hope and intent that her friend will be able to pay off their debt and avoid trouble. She takes a moral gamble, so to speak. Unfortunately, in this sequence of events the gamble does not pay off and the horse loses its race – and so Mary loses all of her friend's money. Mary is intuitively morally responsible for the loss of her friend's money, even though Mary did not intend for this outcome.

When Mary bets her friend's money, her action opens up the possibility of a specific range of outcomes that would otherwise be impossible – the two most obvious possible outcomes being that she loses her friend's money in the gamble or she wins enough money to pay off the friend's debt. As Mary places the bet, she knowingly takes a moral gamble and she is therefore intuitively morally responsible for the outcome of her moral wager, whatever it is. Mary is then, not only responsible for having taken this moral gamble, but specifically responsible for the consequences of it, i.e. the loss of her friend's money.

Let us then return to the notion of manipulation and see where this conclusion about moral gambling takes us. Let us consider a similar scenario but now involving an indoctrinated agent. Suppose that our cult leader believes himself to have knowledge of a guaranteed bet at the races, just like Mary in our previous example. The cult leader wants to ensure such a bet is made, for which the winnings will go straight into his own pocket. Now, suppose the cult leader has taken a special interest in one particular cult follower and, over time, the leader has indoctrinated him to a frightful standard. This particular indoctrinated agent happens to have a large amount of life

savings, and the cult leader eyes an opportunity. As a result of the cult leader's manipulation, the indoctrinated agent bets everything he owns at the races, just as the cult leader wants.

Recall my analysis of what is going on in such cases. The manipulated agent has been deliberately limited by his manipulator in his access to a range of sufficient reasons to do otherwise. So there might be a long list of reasons why our agent should not bet all his life savings on a horse; for example, suppose that in some nearby possible worlds his mother is very ill and he should send money to her. However, our indoctrinated agent is not receptive to these reasons: in these nearby possible worlds, the manipulator sees to it that the agent does not become aware of the excellent reasons not to make the bet. To repeat, however, the agent is receptive to *some* reasons to do otherwise. For example, in nearby possible worlds where the cult leader decides that the bet is too risky, they will ensure that the agent is aware of *that* reason not to make the bet.

Now, just as in the case of Mary, the wrong horse wins and all the money is lost. The cult leader, through his indoctrination of the cult follower, orchestrated this bet coming about, however the cult leader obviously did not want or intend this outcome. Nevertheless, his indoctrination of this agent in the given scenario opens up a specific range of possible outcomes that would otherwise not come about; two of these outcomes of interest here, of course, being that either all the money is lost or that the gamble pays off. The cult leader is therefore making a moral gamble and, even though he might have not wished or even intended for this scenario to happen, he still decided to roll the dice. As such, he is morally responsible for the outcome of the gamble, whatever it turns out to be, just as much as he would be if he could predict exactly what the outcome would be.

Manipulators are morally responsible for the events caused by their indoctrination, then, even if they do not have full foresight about them. In the following section, I will seek to

identify the manipulators of autonomous systems and use the conclusions of the previous sections to solve the problem posed at the start of this thesis by identifying who is morally responsible, and why, in contemporary cases featuring moral responsibility gaps.

IV. The Manipulators of Autonomous Systems

Autonomous systems are examples of peculiar technology, unsettlingly futuristic in both their presentation and use. I use the word 'peculiar' here because autonomous systems are both creations as well as potential agents in their own right. I showed in Chapter 5 that autonomous systems easily fulfil the conditions for moderate reasons-responsiveness, which makes it inappropriate to classify them as mere tools. Nonetheless, as I stated at the start of Chapters 2 and 3, there may be a wide range of reasons to think that autonomous systems cannot be morally responsible. After all, I have only discussed the freedom-relevant conditions for moral responsibility, and I have therefore left it open that there are epistemic conditions for moral responsibility that these systems do not, and perhaps cannot in the near future, fulfil. Hence, full attribution of moral responsibility to these systems has never been on the table in this thesis. Moreover, in §II of this chapter, I argued that contemporary and near-future autonomous systems, while being moderately reasons-responsive, fails to have guidance control, as their creation by humans for limited purposes relegates them to the category of manipulated agents. However, it is this categorisation of autonomous systems as manipulated entities that, I will argue, can be used to shed light on the moral responsibility gap problems presented back in Chapter 1. In this section I will argue that the users of autonomous systems are functionally analogous to manipulators of autonomous systems. In the following section, I will then show how these conclusions affect the proper analysis of moral responsibility gap problems and allows us to bridge the moral responsibility gap.

Throughout this thesis, I have been taking seriously the thought of considering autonomous systems to be potential candidates for moral agency. In the course of doing so, I have argued that autonomous systems may fulfil one of the conditions for guidance control, namely that of moderate reasons-responsiveness. The system's moderate reasons-responsiveness explains the intuition that an autonomous system is in a sense in control of its own outputs. Further, the reasons-responsiveness of autonomous systems makes it intuitively clear why their use is creating strange cases; autonomous systems are not just unpredictable tools, but instead entities capable of autonomous reasons-responsive decision-making that can have poor outcomes – and, as with human agents, this need not be the result of a malfunction. Hence, understanding autonomous systems as reasons-responsive entities makes it clear both why moral responsibility gaps come about and why they are so puzzling.

Autonomous systems, then, are a strange kind of 'agent'. There is no doubt that there is a long road ahead before we get to the point where we should consider autonomous systems to be fully morally responsible for their outputs. And yet, these systems are – right now – able to generate independent outputs, seemingly leaving no one directly responsible for them. I have throughout this thesis referred to autonomous systems as 'agents' for ease. Despite my liberal use of the term, however, I have not argued, and will not argue, for the claim that autonomous systems should be conceived as full moral agents. Nevertheless, autonomous systems can and should be seen as reasons-responsive entities. I will show that this is sufficient for solving the original problem of this thesis presented back in Chapter 1: the moral responsibility gap.

As shown in §II, indoctrination of human agents can be analysed as the severe limitation of an agent's reasons-receptivity, usually done to promote the interests of the manipulator. At a first glance, manipulation might seem an odd topic to pursue in relation to

autonomous systems. Historically, we usually think of manipulation as something that happens to human agents. However, insofar as indoctrination is largely defined (following Yaffe's definition) by its effect on the reasons-receptivity of the intended target, there is nothing about indoctrination as a concept that stops us from applying it to autonomous systems. In the simplest of terms, manipulation is just the practice of limiting an agent's behaviour in order to promote one's self-interest. Autonomous systems are entities of limited behaviour/outputs, purposefully created in order to promote the interests of someone else – for example, a company, a government, an individual consumer and so on. Hence, as I showed in §II of this chapter, autonomous systems can be viewed as a type of manipulated reasons-responsive entity. Two questions in particular must follow such a conclusion: (1) who is identifiable as the manipulator of a given autonomous system, and (2) how does this identification affect analyses of moral responsibility gap cases? I will seek to answer the first question here and the second in the following section.

The question at hand, then, is how to identify the manipulator in cases featuring autonomous systems. Here there are only two possible options. Recall from Chapter 1 that only two groups of human beings are at any point directly involved with an autonomous system in use: the creators and the users.[105] I will argue that while both groups can be found to be involved with the manipulation process, ultimately the user(s) of an autonomous system can be found to be the morally responsible party for its outputs.

As was briefly discussed in Chapter 1, autonomous systems and contemporary robots are rarely, if ever, designed and created by just one person. Instead, usually the creation of such a system includes multiple different groups of researchers all working on their own individual aspects

[105] I write 'directly', in order to not include something like pedestrians hit by self-driving cars or other bystanders.

of the system. I will come back to that in a moment. For now, let us focus on the actual process of creating an autonomous system. In the creators' making of the system, they inherently limit it. When creating an autonomous system designed for medical procedures, the system is made receptive to things relating the surgical field. But it will not be made receptive to what the weather is like, what music is played in the operating suite, nor whether the patient is nervous or scared before the surgery. The designers/creators specify the general scope of a system's reasons-receptivity, by giving it its sensors and specifying the general methods of their use. In this way, the creators as a group can be considered partly analogous to an active manipulator, as they form the reasons-responsive system – including its limited reasons-receptivity – for a specific purpose. It may be, of course, that for technological reasons they are simply unable to equip the system with certain kinds of sensitivity that it would be good, in principle, for the system to have. Nonetheless, in the sense explained in the previous section it is still true that the creators have purposefully limited the system's range of reasons-receptivity to the extent that the system counts as a manipulated agent.

I will return to the role of the creators in a moment, but for now let us take a closer look at the role of the users of autonomous systems. The users make up the second category of humans involved with autonomous systems in any given case. Their involvement with autonomous system takes a different form from the creators', though it is fairly self-explanatory. The users … well, they make use of autonomous systems in one way or another.

When using an autonomous system, the user decides such things as the whereabouts of its deployment, the frequency of its use and to what ends the system is being used. As an example, for a surgical system, the user would choose the operating suite or hospital for the system to be deployed in, when to use the system and for what type of surgeries. Similarly, for the cleaning

system example used earlier, the user designates the messy room as the system's available area for functioning in, they choose when to put in the cleaning robot and the purpose of its deployment. The creators put these entities into the world, however the users activate them and decide the when, where and the purpose of their use. As the user sets a task or releases an autonomous system in their chosen environment, the user utilises a system with limited reasons-receptivity, usually to promote their own interests or the interests of their fellow humans. In this way, independent of the reasons that the users may have to use a given autonomous system, the user is involved in the process that leads autonomous systems to provide their outputs.

Now, the roles of the creators and the users seem clear, however still neither of them look like obvious 'manipulators' in the sense that the cult leader from §II and §III was. I will argue that this can explained by showing that the creators and users together make up the manipulator role in cases featuring autonomous systems. Usually, manipulation cases used in philosophical thought experiments features a single manipulator (such as the cult leader), but a 'divided manipulator' role is fairly easily imagined. For example, imagine a society similar to the one described by Aldous Huxley in *Brave New World* (1932). A key attribute of this society is how the society uses indoctrination to secure peace. Citizens are made with the use of artificial wombs, and through the use of lengthy childhood indoctrination programmes, they are divided into different castes based on intelligence and use.

Imagine that an agent is made in a reproduction centre in this indoctrination-heavy society. At the reproduction centre, the creators of the agent limit the agent's reasons-receptivity in a way that is conducive to his designated purpose in society.[106] Suppose that the limitations are

[106] Outside of classic psychological indoctrination, the lower castes in Brave New World are also victims of serious physical limitations, which in the novel is done by altering the growing environment for the agents, when they are still developing foetuses. This particular limitation of the physical aspects of an

made to make the agent a soldier with certain qualities that are revered in this society. Specifically, the agent is limited in his reasons-receptivity. As such, he will not be receptive to a wide range of reasons that include things such as empathy, fear for one's own life and similar. In practise, this means the agent will not form any emotional connections to other people, thereby eliminating the possibility of anyone mourning him if he were to fall in battle. Further, he will not be in danger of developing any long-lasting mental problems, such as post-traumatic stress disorder or depression following combat. Along with blind obedience, the agent has the psychological attributes – or rather, limitations – to become this society's version of the perfect soldier. The first element of the soldier's manipulation is thus easy to identify: it is found in his very creation and in his formative years influenced by his creators in the recreation centre.

Next, consider our soldier being utilised in this society. If he is in the military, he will have a commander setting his tasks every day. Due to the limits imposed by his indoctrination, the soldier is only receptive to the reasons that his commander has for having him perform behaviour in a certain way. As such, our soldier spends day after day doing his commander's bidding without questioning thanks to his deep-running indoctrination. The commander here plays the second manipulator role.

The commander did not do the basic indoctrination of the soldier himself. However, the commander knows that the soldier is of an indoctrinated kind, and chooses to utilise the fact that the soldier is in this state to further his own interests. As such, two roles in the manipulation case exists. First, there is the person or team of people who do the actual indoctrination (including the manipulation of the environment in which the embryo develops), ensuring that the agent in question

agent is something that can be seen not only in the human agents in Brave New World, but also in autonomous systems in our actual society.

is limited in a certain way. This role is in this case played by the creators at the reproduction centre. Second, there is the one who uses the indoctrinated agent to further their own agenda. Here this role is played by the commander. In classic thought experiments involving manipulation, the two roles of manipulation are usually played by either the same agent, or there is only one level of manipulation in play. Thus, for example, in Mele's (2006) 'zygote argument', the manipulator, Diana, wishes for an event E to happen. She creates a zygote that is determined to end up as a human agent, Ernie, who will perform an action A, thereby leading to the specific event, E, happening. Yet, once Diana has created the zygote, she does not intervene in any way (Mele 2006, 188). However, as shown by the 'brave new world' example, there can be distinct manipulators with largely independent roles, whose actions together determine the manipulated agent to act as they do on a specific occasion.

Now, one might wonder why I created an example that drew inspiration from dystopian fiction. The answer to that is simple. In general society, human agents are neither designed nor indoctrinated in such a systematic way as to make it possible to invent a thought experiment that would be at all closely analogous to scenarios featuring manipulated autonomous systems. The thought experiment above is therefore, thankfully, largely fictional. Nevertheless, despite the fictional state of the society, the story still shows that intuitively one can imagine the manipulation in the creation and development of a human agent and the use of that manipulation to be performed by two independent agents/groups of agents. In this way, accepting the two-part manipulation of autonomous systems in our world should become easier.

As such, two independent agents/groups of agents can together fulfil the manipulator role. The creators and users of autonomous systems play analogous roles to, respectively, the indoctrination team and the commander in the *Brave New World* example set out above. In this

way, the creators and users together can fulfil the manipulator role in relation to a limited reasons-responsive autonomous system.

Now, I have argued that both users and creators of autonomous systems each play a role in the manipulation process of these systems, analogous to the roles played by the people at the reproductive centre and the commanding officer in the thought experiment above. I will now argue that due to their different roles in the manipulation process, their moral responsibility vis-à-vis the outputs of the autonomous systems also differs. Let us start by returning to the discussion of the role of the creator.

I will start by loosely recounting an incident I witnessed some years ago.[107] I had been invited to an ethics meeting for a research group that was discussing selling the patent for one of their creations. The sale in question would be to a private firm in a country, whose governmental branches has a long history of infringing on the rights of its citizens. To the best of the research group's knowledge, if sold to this company the product would end up being used to track the citizens of the country in question.

The PI argued for selling, and he made the following analogy. He said that what they had made was like a paperclip, nothing more. A paperclip can be used for collating papers, or for pushing under a wriggling table, or it can be bent and used to stab someone. In his mind, they had made something akin to a paperclip, and the uses to which someone might put their paperclip was not on them.

The same move might be made by the creators of autonomous systems. They have created this entity and opened up the possibility of its use. But they are not forcing anyone to use it,

[107] I have here altered some details of this story for obvious reasons.

nor are they dictating in what circumstances their creation should be used, or to what purpose, and so they bear no responsibility when it comes to autonomous systems with outputs that have hurt human beings. It is clear, however, that this kind of defence seems rather contrived, especially when considering similar scenarios. Imagine that a group of creators have created a death-inator, a type of weapon that sends out a ray killing people in an instant. Suppose they sell it to a stranger, who turns out to be a truly heinous villain, who uses the invention to purposefully murder thousands of innocents. In such a case, it seems highly counter-intuitive to declare the creators free of any moral culpability in relation to the death-inator and its victims. As such, the responsibility of creators cannot so easily be dismissed.

Creators of autonomous systems are morally responsible for the creation of their inventions. As such, an argument may be made that creators are morally responsible for putting these entities into the world, and potentially for carelessness regarding their creation's misuse. As we discuss the possible responsibility of creators, I must remind the reader that for the purposes of this thesis we are dealing with non-malfunctioning autonomous systems only. Of course, in cases of malfunction, particularly due to neglect during the creation process, it is only natural to look back to creators as the responsible party.

However, when it comes to non-malfunctioning autonomous systems, I shall refer to the creators' intuitive responsibility for putting their creation into world as a general responsibility vis-à-vis the creation of autonomous systems. However, recall from Chapter 1, §II and §III, that this thesis is not interested in purely some general responsibility in relation to autonomous systems – instead, we are looking for a party responsible for the specific outputs of these systems and the harms that may arise directly from these. Hence, it is useful to distinguish between these two types of responsibilities:

1. Responsibility for creation – this responsibility being tied with the general responsibility that might arise from putting a certain system into the world and its future development.
2. Responsibility for a specific output and its consequences.

That there is a distinction between these two should be fairly obvious, as it mimics the discussions we had all the way back in Chapter 1, §II and §III, in relation to the current literature on creator and user responsibility. However, I will argue that placing the first type of responsibility onto creators is not equal to placing the second type on them. Hence, despite the role they play in the manipulation process or creation of autonomous systems as entities with limited reasons-receptivity, the creators are still not clearly morally responsible for the systems' individual outputs.

The appropriateness of the first type of responsibility relies intuitively on the type of autonomous system which is brought into the world as well as the intentions surrounding its creation. In the 'death-inator' example, it seems completely fair to say that the creators are morally responsible to some extent for the deaths of the civilians, as they created a thing that was designed to kill and put it into the world without much control of who would use it. In contrast, consider a team creating a cleaning system. Even if somebody were to misuse the system – say asking it to mop a specific area, thereby locking inside a hydrophobic family member – it would be strange to say that the creator team was morally responsible for this. After all, they just created a cleaning system and, in this way, the case seems analogous to the 'paperclip creation' example. Hence, when it comes to creator responsibility in the first sense, it seems that this type of responsibility depends on the nature of the autonomous systems in question and the specific purposes for which it was designed.

In contrast, the second type of responsibility is tied to the specific output of a given autonomous system. In other words, to be responsible for the output is to be responsible for the autonomous system's specific reasons-responsive output/action and its immediate consequences. Such a responsibility is tied to the output and therefore is independent of both the nature of the given autonomous system and the specific purposes for which it was designed. In other words, the responsibility that we are trying to place here – responsibility for a system's output – is not dependent on whether the system being talked about is a cleaning system or one made for war, but instead tied to the output being a manipulated reasons-responsive output/action. Insofar as we can not tie the creators' role in the manipulation process directly to the individual outputs, then we can not put the responsibility for specific outputs on their table. As such, while creators may have some kind of general responsibility vis-à-vis the creation of these systems, they cannot be directly tied to the individual outputs and as such we cannot place the type of responsibility on them which could help us solve moral responsibility gap problems.

Instead, I will argue that the user's role in the manipulation process of an autonomous system makes it possible to hold them morally responsible for the outputs in moral responsibility gap cases. To see why, let us again return to our Huxley-inspired world. Suppose that a private person makes use of the cold efficiency of the 'designed' soldiers. More specifically, imagine that he tasks these indoctrinated soldiers to stand guard around his factory, as lately he has had trouble with the workers striking and threatening his profits. The factory owner knows that the instatement of the soldiers on the factory floor will probably either frighten the workers back to the assembly line, or, in the case of a violent rebellion, protect his assets by violent means. The soldiers do not feel empathy with the factory workers, nor will they rebel themselves in solidarity thanks to the indoctrination. One day, the workers rebel and it ends in fighting, resulting in multiple casualties.

Now, the factory owner has nothing to do with the original indoctrination – as such, he does not possess the creator role with respect to these indoctrinated soldiers. Yet, the factory owner is the one making use of the indoctrinated agents and their naturally limited state. He puts them in a scenario where he can reasonably expect certain outcomes based on the indoctrinated agents' lack of access to the range of reasons to do otherwise. He thus uses the indoctrinated agents as tools to ensure as best he can that the workers are held down in one way or the other. Just as we hold a user morally responsible for using a tool resulting in harm to others, so do we hold manipulators or users of indoctrinated agents morally responsible for their use resulting in harm. Of course, this is not to absolve the people who have created these soldiers fully from any moral culpability – there is still space for that, as mentioned earlier in this section. But it does mean that just as the factory owner is morally responsible for the deaths of his workers, the users of indoctrinated agents are morally responsible for the consequences that come about from their use.

It should be clear how this conclusion relates to cases featuring autonomous systems. If a non-malfunctioning autonomous system can be classified as an entity with limited reasons-receptivity, or in other words a type of indoctrinated entity, and its users are the ones who make use of it (like the commander from the *Brave New World* example, or even the cult leader example), then even though the users did not 'indoctrinate' or limit these systems' reasons-receptivity, they can still be considered morally responsible for any outputs of the system that results in harm in virtue of being the party that makes use of the indoctrinated/limited entity.

V. Bridging the Moral Responsibility Gap

Let us now bring back some of the possible moral responsibility gap cases discussed back in Chapter 1. The original thesis of the moral responsibility gap proposes that there are possible cases,

where the output of a non-malfunctioning autonomous system causes harm to a human person, yet nobody appears to be morally responsible for this harm. As shown by writers such as Gunkel (2017), in the last decade such cases have stopped being mere hypotheticals, but instead are actual cases growing in number every year. One such potential case may feature an autonomous system used for self-driving cars.

Imagine the following scenario. A human agent, Amelia, gets in an autonomous vehicle that is level 5 on the scale of autonomy presented in Chapter 5.[108] In practice, this means the vehicle needs no human supervision to function, nor does it need a human to be a 'stand-by driver' for potential hazardous situations. She says/enters her destination, which is registered fine by the system. The system calculates a path and gets going. While en route, an unfortunate chain of events happens such that the system is faced with a dilemma. On one side of the car is a motorcycle and on the other side is a small car containing a family. The self-driving car cannot avoid crashing into one of them and has to make the split-second decision which one to hit. Suppose that in this sequence of events, the system's decision-making results in the car crashing into the motorcyclist. Amelia is, of course, shaken, but otherwise alright, while the motorcyclist has suffered multiple grave injuries.

This is a classic example of the moral responsibility gap. However, five chapters have now passed and we can now bring some new insight to this case. We now know that an autonomous system such as this car may be categorised as a limited reasons-responsive entity, with Amelia as the user of the system. Traffic collisions happen, and this fact is also well-known by any adult agent who uses or drives cars (or similar vehicles). In our above scenario, Amelia chooses to use the self-driving car and does so with the intention that it will bring her to a particular address and, of course,

[108] Note that at the time of writing, self-driving cars at this level of autonomy have not yet been developed nor made commercially available.

not with the intention that the car will be involved in a traffic collision. Nevertheless, her choice constitutes a moral gamble.

Amelia's choice to use the self-driving car is just like the case with Mary and the horse race betting scenario described in §II of this chapter, just with different odds. Recall that Mary in that example was found to be morally responsible for losing her friend's money after betting it on a horse race, even though Mary was convinced that it was a sure-fire bet. While the odds here may be strongly in Amelia's favour, as self-driving cars should on average be safer than the average driver, the gamble does not pay off in this sequence of events. The motorcyclist is harmed. Amelia used a limited reasons-receptive entity and utilised it in an environment of her choosing. She chose to use a limited entity – which, I've argued is a kind of indoctrinated agent. As such, taking on our lessons on moral tracing from the factory owner case as well as the horse race betting case, we see that Amelia is morally responsible for the motorcyclist's injuries.

Consider another classic moral responsibility gap case, this time featuring a military drone. Recall from Chapter 5, as well as back in Chapter 1, that the literature has seen its fill of nightmare scenarios, where this technology in theory could spawn death and destruction with no one morally culpable for the carnage. While the technology is currently a fair way away from fitting this description, let us nevertheless consider the moral consequences of this kind of hypothetical scenario.

Imagine a near-future military squad commander who is dividing assignments between his soldiers, both human and machine. Among these is an autonomous military drone with the capability to fire missiles. It is asked to patrol an area of enemy territory from the sky, while its human colleagues are fulfilling some orders on the ground. The drone calculates an efficient route and gets started. While on duty, the drone unexpectedly observes a high-profile target disappear into

a building. Suppose that the target famously seems to vanish into thin air after each sighting, so there is a termination order on him. Suppose the drone therefore makes the decision to fire. A missile enters the building killing the target and nine innocent civilians in an instant.

When the commander decides to use the drone and give it an assignment, she chooses to make use a of an entity that is severely limited in its range of reasons-receptivity. As such, the commander chooses to make use of an indoctrinated/limited entity in order to further her agenda. In this way, we can see that the commander takes on the second kind of manipulator role: that of a user of the autonomous system. While the commander never had anything to do with the system's actual creation, she was still the one making use of the limited entity, as per the discussion earlier in §IV. Hence, as with the earlier cult leader example, the *Brave New World* example and the factory owner example, the user is morally responsible for the outputs of the indoctrinated agent – here the user being the military commander.

Now, one might say that the commander might have issued the order to the drone intending for it to merely observe and be able to report if any enemies close in. The area where she asked the drone to patrol might even be famously peaceful, so she had no reason to expect the target to be present there. Nevertheless, the commander made a moral gamble by using the system. When she sent in a drone with missile-firing capabilities and information relating to the target into a specific territory, she may have intended and wanted for the drone to sit back and merely observe. But she also knew there was a chance, however small, that the drone could harm people, more specifically civilians, in its line of duty. As such, uncertainty about the system's outputs does not absolve the commander of her moral responsibility. In this way, the example is much like the case of Mary and the horse race betting.

As such, we have a model for the analysis of moral responsibility gap cases. In such cases, if the reasons-responsive autonomous system involved is seriously limited in its range of reasons-receptivity, then the moral responsibility for the system's output befalls the agent who is identifiable as having made use of the system and its limited state. In other words, in moral responsibility gap cases, the moral responsibility for the outputs of a given non-malfunctioning autonomous system can be traced to the system's user. As both contemporary and near-future autonomous systems fall into the category of reasons-responsive entities with a limited range of reasons-receptivity, the above conclusion will apply to the large majority of imaginable moral responsibility gap cases. Hence, the gap can be closed.

In this chapter, I have argued that autonomous systems' limited range of reasons-receptivity matches that of indoctrinated agents, and they are therefore, similarly, a type of manipulated reasons-responsive entities. And I have shown that for actions or outputs born out of this type of manipulation, the moral responsibility lands with the agent who makes use of the agent's limited state. Further, I have argued that this placement of moral responsibility does not shift, even if the agent in question did not have full foresight about what would happen. Hence, in manipulation cases featuring autonomous systems, the moral responsibility for their outputs can be traced to the system's user, thus closing the moral responsibility gap.

Concluding Remarks

In this thesis, I have argued that a solution is possible for moral responsibility gap problems. As part of this solution, I investigated the nature of autonomous systems in order to clarify their potential for fulfilling the freedom-relevant conditions of moral responsibility. I argued for autonomous systems to be recognised not, as mere tools but instead as manipulated reasons-responsive entities. Last, but not least, I concluded that this perspective on autonomous systems, allows us to see their users as morally responsible for the system's outputs and the consequences thereof. As such, this thesis has presented one possible solution to moral responsibility gaps by developing a reasons-responsive machine compatibilist account and explaining why, given that account, it is the users of autonomous systems who are morally responsible when a non-malfunctioning autonomous system produces an unwelcome output.

Yet, this is hardly the whole story. Every chapter of this thesis has discussed or investigated uncharted territories when it comes to research on moral responsibility gaps or, more generally, the moral issues raised by the use of non-malfunctioning autonomous systems, and the thesis as a whole therefore only represents a very narrow investigative path. I will therefore conclude the thesis by briefly considering some of the roads not taken.

In Chapter 1, I introduced Mattias' (2004) notion of moral responsibility gaps. I argued that these present a philosophical problem: when non-malfunctioning autonomous systems provide an output that harms human beings, there is seemingly no one intuitively morally responsible for the harm in question. I showed in this chapter that the classic instrumentalist view of autonomous systems had led to discussions of user and creator responsibility, though in general the writings on this topic still lacked in philosophical rigour and finesse. Hence, though I chose here to disengage with the instrumentalist approach, it must be clear that there is still plenty to say on that account. While we are used to seeing autonomous systems and other machines as just instruments,

Chapter 1 should at least have made it clear that autonomous systems push the boundary for what can reasonable be defined as a mere tool. Autonomous systems, in virtue of being reasons-responsive entities, are at least in this way more like genuine agents than they are like tools as traditionally conceived. As such, if one were to create an instrumentalist account of autonomous systems, their status as mere tools cannot be hastily assumed. Such an account would need to address more carefully how much autonomy an entity can have (if any), while still being considered a tool and nothing more. In truth, even if we relegate autonomous systems to the category of mere tools, they stretch the definition and are therefore naturally philosophically curious.

In Chapter 1, I also showed that previous attempts to investigate creator and user responsibility have yielded few convincing results with regard to bridging the moral responsibility gap. This is partly why I started this thesis by investigating autonomous systems in moral responsibility gap cases without taking on an instrumentalist stance to these systems. Now, in Chapter 6, I ended up concluding that in moral responsibility gap cases featuring current and near-future autonomous systems, it is the users who are morally responsible for the systems' output and the immediate harm that may follow from this output. As such, it is fair to question what the purpose of investigating the possibility of morally responsible systems was, considering that I ended up pointing the moral finger at a human party anyway. The choice to investigate autonomous systems as potential moral agents – and not as just tools – yielded two pivotal findings. The first was the finding that these systems are manipulated reasons-responsive entities. As such, the users could be found morally responsible with respect to a given autonomous system's output, precisely *in virtue* of the system's status as a potential moral entity. The manipulation relationship explains *why* the user is morally responsible for a given output in a clear-cut way that could not be done if an instrumentalist approach had been taken to autonomous systems in general. Hence, the changed perspective on autonomous systems does not merely identify the morally responsible party in moral

responsibility gap cases; it explains why they are morally responsible in a way that takes account of their relationship to the system itself, which in turn takes account of the fact that the system has a degree of autonomy that is not found in more traditional tools such as hammers and laptops.

The second finding was that autonomous systems, while not fully fledged moral agents, were shown to be reasons-responsive entities. Even though I concluded in this thesis that current and near-future autonomous systems would qualify as manipulated entities, thereby relegating the moral responsibility to their manipulator making use of them – my machine compatibilist account and analysis of moral responsibility gap cases leaves it open that future autonomous systems could one day potentially fulfil the conditions for the freedom-relevant condition for moral responsibility. As such, by considering the possibility of morally responsible autonomous systems, I have created an account that not only sheds light on current and near-future autonomous systems, but can also be used for analysis of future autonomous systems, as the field of autonomous systems develops and such systems become increasingly sophisticated.

By not using the instrumentalist perspective on autonomous systems, the machine compatibilist account presented in this thesis allows for further development and changes to the technology while still remaining relevant as a solution to moral responsibility gap cases. If and when the day arrives where autonomous systems fulfil the freedom-relevant condition for moral responsibility set out in this thesis, however, a further question will arise concerning what else is needed – aside from the freedom-relevant conditions my account borrows from Fischer and Ravizza (1998) – for these systems to be considered moral agents. Such further questions have been left open in this thesis. In short, my account has identified a necessary condition on moral responsibility that current and near-future autonomous systems fail to fulfil; at the point where autonomous

systems fulfil that condition, it will need to be considered whether or not there are other conditions that they still fail to fulfil.

In Chapter 2, I presented an argument for the 'machine incompatibilist' stance, according to which autonomous systems – being deterministic – cannot be morally responsible, and considered some potential traditional compatibilist ways to counter this line of argument. The investigative choices made in this chapter are easy to trace: I focused on the freedom-relevant condition of moral responsibility and identified Bringsjord and similar writers as machine incompatibilists, who appeal to a machine-specific version of the Consequence Argument, which in turn naturally lead to considering 'machine compatibilist' responses to the argument.

While my account firmly belongs in the 'machine compatibilist' category, both machine compatibilism and machine incompatibilism represent significant new possible research areas. While I aimed for providing a rebuttal to the machine incompatibilists' machine-focused version of the Consequence Argument, one could just as easily pursue other machine incompatibilist routes taking inspiration from the arguments of, for example, Kane (1989), G. Strawson (1986 and 1994), and Pereboom (2001). Similarly, while my machine compatibilist account was created with current and near-future autonomous systems explicitly in mind, there is nothing to stand in the way of developing machine compatibilist accounts for their own sake. Without the requirement of the account providing a necessary positive story of autonomous systems and moral responsibility, such accounts could instead shed light on what it exactly it would take for autonomous systems to fulfil different freedom-relevant conditions for moral responsibility set by a variety of established compatibilist theories. I'll return in a moment to discuss why this might be a worthy pursuit in itself.

So, it should be fairly easy to imagine how vast the potential research area focusing on autonomous systems and the freedom-relevant condition for moral responsibility is, not to mention that work on epistemic conditions for moral responsibility has shown that these also provide their own challenges in relation to attribution of moral responsibility, independently of freedom-related worries. Following in the footsteps of writers such as Fischer and Tognazzini (2009) and Levy (2011 and 2014), there is therefore also plenty of space to develop a sceptical (or perhaps a positive) account of how and why autonomous systems might fulfil, or fail to fulfil, various epistemic conditions on moral responsibility.

In Chapter 3 and 4, I investigated the possibility of using the accounts of Wolf (1987), Strawson (1962) and Fischer and Ravizza (1998) as the foundation for my own machine compatibilist account. I chose not to develop a machine compatibilist account based on either Wolf or Strawson's theories, as I came to the conclusion that such accounts would have little of interest to say in relation to current and near-future autonomous systems' roles in moral responsibility gaps. It is, however, fairly obvious that without a special focus on moral responsibility gaps or even current/near-future technology, it is hypothetically possible to investigate any traditional compatibilist account of moral responsibility for its possible extension to cover non-human agents and entities such as autonomous systems. While such pursuits might just be done out of curiosity or as an intellectual exercise, such projects also carry some practical merit. In short, considering the viability of extending any given compatibilist account to cover autonomous systems or other decidedly non-human entities can help highlight potential problematic assumptions or speciesist requirements for moral responsibility. For example, I argued in Chapter 3 that Strawson makes use of the 'reversal thesis' which inherently seems to keep out non-human agents or lesser sophisticated agents out of any potential contention for moral responsibility. By considering the possibility of a Strawsonian machine compatibilist account, it was revealed that such an account would not allow

for less-than-human agents to be considered part of our moral community, eliminating the possibility of them ever being considered morally responsible. Similarly, in Chapter 4, I mentioned that it has been previously argued that Fischer and Ravizza's reasons-responsive account of moral responsibility does not need the mechanism ownership aspect to work. However, without this criterion or some kind of no-manipulation clause (as was discussed in Chapter 6), even the simplest autonomous systems (such as the turtles introduced in Chapter 5) can be ascribed guidance control and thereby fulfil the freedom-relevant condition for moral responsibility. Yet, such ascriptions would strike most people as intuitively wrong and such a conclusion would therefore naturally call for a re-evaluation of the account.

As such, when creating machine-specific versions of classic compatibilist accounts it becomes more than just some intellectual endeavour. Autonomous systems or similar robotic systems, when considered in conjunction with a compatibilist account, can be used as hypothetical test-dummies. They are entities with no or little human features, but instead only the most basic capabilities for producing action/outputs. Hence, when considering them in relation to different criteria for moral responsibility, we cut to the bone of the conditions set up by these accounts and get a clearer insight into whether these conditions are truly enough to capture the conditions for moral responsibility. Hence, machine compatibilism can provide new insight into established compatibilist accounts and pinpoint their potential weaknesses.

In Chapter 5 and 6, I argued that autonomous systems are manipulated reasons-responsive entities. As such, it was in these chapters that I argued for a specific account within the field of machine compatibilism. In order to create this account, I make some potentially controversial argumentative moves, including using Dennett's (1978) intentional stance to ascribe reason-based decision-making to autonomous systems, extending Yaffe's (2003) definition of

indoctrination to autonomous systems in order to describe these as manipulated entities, and so on. I made these moves in order to force a positive story of autonomous systems and the freedom-related conditions for moral responsibility. I am well aware that someone who is sceptical about my account has plenty of ammunition, considering that some of these moves are rather unorthodox. Nevertheless, I shall briefly try to show that any scepticism towards my account does not diminish the value of the work done in this thesis.

If a sceptic rejects any part of my thesis, then, the whole account does not tumble down like a house of cards. Instead, the sceptic is left with the outlines of multiple new research topics, as has been shown in these concluding remarks. If one rejects the idea that autonomous systems are manipulated entities, for example, one is still left with the idea that they are reasons-responsive entities, which is still a new perspective on autonomous systems. If one rejects the idea of using Fischer and Ravizza's theory as the foundation for a machine compatibilist account, then one is still left with many other possible ways of developing machine compatibilism. If one is sceptical about machine compatibilism in general, then machine incompatibilism presents itself as an area of interest – but one that requires more argument in its favour than merely the machine-focused version of the Consequence Argument discussed in Chapter 2. And so on. In this thesis, I have not presented an account that is situated within a well-established field; instead, I have carved out the specific account as well as made a rough guide for the field it lies within. My real aim was not so much to develop, *ex nihilo*, a watertight version of machine compatibilism. It was rather to spark some life into the discussion surrounding moral responsibility gaps by taking one particular machine compatibilist path, while pointing towards the many other possible paths one could travel down. Problems with the account I have presented constitute pointers towards other – perhaps better – paths that could be taken.

Moving forward, I therefore hope that my thesis may serve as both an example of, and a roadmap for, the possibilities for philosophical research involving autonomous systems and moral responsibility.

Bibliography:

Allen, C., Varner, G., and Zinser, J. 2000. 'Prolegomena to any Future Artificial Moral Agent', *Journal of Experimental & Theoretical Artificial Intelligence,* 12(3): 251-261.

Allen, M.J., and Edwards, I. 2021. *Criminal Law*. Sixteenth edition. Oxford: Oxford University Press.

Arbib, M.A., and Fellous, J.M. 2004. 'Emotions: from Brain to Robot', *Trends in Cognitive Sciences*, 8 (12): 554–561.

Aristotle. 1985. Nicomachean Ethics (Trans. by T. Irwin). Indianapolis: Hackett.

Arkin, R.C. 2009. *Governing Lethal Behaviour in Autonomous Robots*. London: Chapman and Hall/CRC.

— 2010. 'The Case for Ethical Autonomy in Unmanned Systems', *Journal of Military Ethics*, 9(4): 332–341.

—, Ulam, P., and Wagner, A.R.. 2012. 'Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception', *Proceedings of the IEEE*, 100 (3): 571-589.

Ashrafian, H. 2017. 'Can Artificial Intelligences Suffer from Mental Illness? A Philosophical Matter to Consider', *Science and Engineering Ethics*, 23: 403–412.

Attanasio, A., Scaglioni, B., De Momi, E., Fiorini, P., Valdastri, P. 2021. 'Autonomy in Surgical Robotics', *Annual Review of Control, Robotics, and Autonomous Systems,* 4 (1): 651–679.

Ayer, A. J. 1954. ‘Freedom and Necessity’, in his *Philosophical Essays*. London: Macmillan.

Bartneck, C., and Hue, J. 2008. ‘Exploring the Abuse of Robots’, *Interaction Studies,* 9 (3): 415-433.

Beebee, H. 2003. ‘Local Miracle Compatibilism’, *Noûs,* 37: 258-277.

Beer, J.M., Fisk, A.D., and Rogers, W.A. 2014. ‘Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction’, *Journal of Human-Robot Interaction,* 3 (2): 74-99.

Beiker, S. 2012. ‘Legal Aspects of Autonomous Driving’, *Santa Clara Review,* 52 (4): 1145-1156.

Bennett, J. 1980. ‘Accountability’, in Z. van Straaten (ed.) *Philosophical Subjects: Essays Presented to P.F. Strawson.* Oxford: Clarendon Press.

Bilgrami, A. 2006. *Self-Knowledge and Resentment*. Cambridge, MA: Harvard University Press.

Birdsall, M. 2014. ’Google and ITE : the road ahead for self-driving cars’, *ITE Journal,* 84(5): 36–39.

Block, N. 1978. ‘Troubles with Functionalism’, *Minnesota Studies in the Philosophy of Science,* 9: 261-325.

Bratman, M. 2007. *Structures of Agency: Essays*. New York: Oxford University Press.

Breazeal, C. 2002. *Designing Sociable Robots*. Cambridge, MA: MIT Press.

— 2003. ‘Towards Sociable Robots’, *Robotics and Autonomous Systems,* 42: 167-175.

— 2009. ‘Role of Expressive Behaviour for Robots that Learn from People’, *Philosophical Transactions of the Royal Society B,* 364: 3527–3538.

—, Siegel, M. et al. 2008. ‘Mobile, Dexterous, Social Robots for Mobile Manipulation and Human-Robot Interaction’, in the conference proceedings for *SIGGRAPH’08: Special Interest Group on Computer Graphics and Interactive Techniques Conference*. Los Angeles, California: 11-15 August.

Brey, P. 2013. ‘From Moral Agents to Moral Factors: The Structural Ethics Approach’, in P. Kroes and P.-P. Verbeer (eds.) *The Moral Status of Artifacts.* Dordrecht: Springer.

Bringsjord, S. 2008. ‘Ethical Robots: the Future Can Heed Us’, *AI & Society*, 22: 539–550.

—, and Govindarajulu, N.S. 2019. ‘Artificial Intelligence’, in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, (Winter 2019 Edition). URL = <https://plato.stanford.edu/archives/win2019/entries/artificial-intelligence/>.

—, Xiao, H. 2000. ‘A Refutation of Penrose’s Gödelian Case Against Artificial Intelligence’, *Journal of Experimental and Theoretical Artificial Intelligence*, 12: 307–329.

Brink, D., and Nelkin, D. 2013. ‘Fairness and the Architecture of Responsibility’, in D. Shoemaker (ed.) *Oxford Studies in Agency and Responsibility, Volume 1*. Oxford: Oxford University Press.

Burge, T. 1996. ‘Our Entitlement to Self-Knowledge’, *Proceedings of the Aristotelian Society*, 96: 91–116.

Campbell, J.K. 2005. ‘Compatibilist Alternatives’, *Canadian Journal of Philosophy*, 35 (3): 387-406.

Carlson, E. 2000, ‘Incompatibilism and the Transfer of Power Necessity’, *Noûs*, 34(2): 277–290.

Caruso, G. 2021. ‘Skepticism about Moral Responsibility’, in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), URL = https://plato-stanford-edu.manchester.idm.oclc.org/archives/sum2021/entries/skepticism-moral-responsibility/

Chalmers, D. 1996. *The Conscious Mind.* Oxford: Oxford University Press.

Champagne, M. and Tonkens, R. 2015. ‘Bridging the Responsibility Gap in Automated Warfare’, *Philosophy and Technology*, 28 (1): 125-137.

Chisholm, R. M. 1957. *Perceiving*. Ithaca: Cornell.

— 1964 (1982). ‘Human Freedom and the Self’, in G. Watson (ed.) *Free Will*. New York: Oxford University Press.

Chomsky, N. 1959. ‘Review of Verbal Behavior’, *Language*, 35: 26–58.

— 1971. ‘The Case Against B. F. Skinner’, *New York Review of Books*, 30: 18–24.

Christman, J. 1991. ‘Autonomy and Personal History’, *Canadian Journal of Philosophy*, 21 (1): 1–24.

— 2018. "Autonomy in Moral and Political Philosophy", in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Spring 2018 Edition)*, URL = <https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/>.

Clarke, A.C. 1968. *2001: A Space Odyssey.* London, UK: Hutchinson

Coates, D.J. and Tognazzini, N.A. 2012. ‘The Contours of Blame’ in D.J. Coates and N.A. Tognazzini (eds.) *Blame: its Nature and Norms*. Oxford: Oxford University Press.

Coeckelbergh, M. 2010. ‘Moral Appearances: Emotions, Robots, and Human Morality’, *Ethics and Information Technology*, 12: 235–241.

— 2016. ‘Responsibility and the Moral Phenomenology of Using Self-Driving Cars’, *Applied Artificial Intelligence*, 30 (8): 748-757.

— 2016B. ‘Care Robots and the Future of ICT-mediated Elderly Care: a Response to Doom Scenarios’, *AI & Soc,* 31: 455–462.

Çürüklü, B., Dodig-Crnkovic, G. and Akan, B. 2010. ‘Towards Industrial Robots with Human-Like Moral Responsibilities’, *The 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 85–86.

Danaher, J. 2016. ‘Robots, Law and the Retribution Gap’, *Ethics and Information Technology*, 18: 299-309.

Darwall, S. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Das, S.K. 2008. *Foundations of Decision-Making Agents: Logic, Probability and Modality*. Singapore: World Scientific.

Davenport, J.J. 2002. ‘Fischer and Ravizza on Moral Sanity and Weakness of Will’, *The Journal of Ethics*, 6 (3): 235-259.

Davidson, D. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

— 1991 (2001). ‘Three Varieties of Knowledge’ in his *Subjective, Intersubjective, Objective: Philosophical Essays, Volume 3*. Oxford: Oxford University Press.

Dennett, D.C. 1978. *Brainstorms*. Cambridge, MA: MIT Press.

— 1987. *The Intentional Stance*. Cambridge, Mass.: MIT Press.

— 1991. 'Real patterns', *The Journal of Philosophy*, 88: 27–51.

— 2000. 'With a Little Help from my Friends', in D. Ross, A. Brook, D. Thompson (eds.) *Dennett's Philosophy: a Comprehensive Assessment*. Cambridge, MA: MIT Press.

Dorr, C. 2016. 'Against Counterfactual Miracles', *Philosophical Review*, 125: 241–286.

Duffy, B.R. 2003. 'Antropomorphism and the Social Robot', *Robotics and Autonomous Systems,* 42: 177-190.

Dworkin, G. 1970. 'Acting Freely', *Noûs,* 4: 367–383.

Earman, J. 2007. 'Aspects of Determinism in Modern Physics', in J. Butterfield and J. Earman (eds.) *Philosophy of Physics*. Amsterdam: Elsevier.

Elton, M. 2003. *Daniel Dennett: Reconciling Science and Our Self-Conception*. Oxford: Polity Press.

Enoch, D. 2012. 'Being Responsible, Taking Responsibility, and Penumbral Agency', in U. Heuer and G. Lang (eds.) *Luck, Value, and Commitment: Themes From the Ethics of Bernard Williams*. Oxford: Oxford University Press.

Eronen, M.I. 2020. 'Interventionism for the Intentional Stance: True Believers and Their Brains', *Topoi*, 39: 45-55.

Eshleman, A.S. 2001. 'Being is not Believing: Fischer and Ravizza on Taking Responsibility', *Australasian Journal of Philosophy*, 79 (4): 479-490,

Fara, M. 2008. ‘Masked Abilities and Compatibilism’, *Mind*, 117: 843–865.

Fischer, J. M. 1983. ‘Incompatibilism’, *Philosophical Studies*, 43: 127-137.

— 1986. *Moral Responsibility*. Ithaca: Cornell University Press.

— 1987. ‘Responsiveness and Moral Responsibility’, in F. Schoeman (ed.) *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge: Cambridge University Press.

— 1994. *The Metaphysics of Free Will: An Essay on Control*. Oxford: Wiley Blackwell

— 2004. ‘Responsibility and Manipulation’, *Journal of Ethics*, 8: 145–77.

— 2006. ‘The Free Will Revolution (Continued)’, *Journal of Ethics*, 10: 315–45.

— 2012. *Deep Control: Essays on Free Will and Value*. Oxford: Oxford University Press.

—, Ravizza, M. 1998. *Responsibility and Control: a Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

—, Ravizza, M. 2000. ‘Replies’, *Philosophy and Phenomenological Research*, 61 (2): 467–80.

—, Tognazzini, N.A. 2009. ‘The Truth about Tracing’, *Noûs*, 43 (3): 531-556.

Foley, R. 1979. ‘Compatibilism and Control over the Past’, *Analysis*, 39 (2): 70–74.

Fong, T., Nourbaksh, I. and Dautenhahn, K. 2003. ‘A Survey of Socially Interactive Robots’, *Robotics and Autonomous Systems*, 42: 143-166.

Frankfurt, H.G. 1969. ‘Alternate Possibilities and Moral Responsibility’, *Journal of Philosophy*, 66 (23): 829-839.

— 1971. ‘Freedom of the Will and the Concept of a Person’, *The Journal of Philosophy, 68* (1): 5–20.

Gerdes, A. 2018. ‘Lethal Autonomous Weapon Systems and Responsibility Gaps’, *Philosophy Study*, 8 (5): 231-239.

Gert, B. and Duggan, T.J. 1979. ‘Free Will as the Ability to Will’, *Noûs*, 13 (2): 197-217.

Gertler, B. 2011. *Self-Knowledge*. New York: Taylor & Francis Group.

— 2021. ‘Self-Knowledge’, in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2021). URL = <https://plato.stanford.edu/archives/win2021/entries/self-knowledge/>.

Gibbs, S. 2014. ‘Google’s Self-Driving Car: How does it work and when can we drive one?’, *The Guardian*, 29 May. URL= https://www.theguardian.com/technology/2014/may/28/google-self-driving-car-how-does-it-work (Accessed 9 January 2019).

Ginet, C. 1966. ‘Might We Have No Choice?’, in K. Lehrer (ed.), *Freedom and Determinism*. New York: Random House.

— 1980. ‘The Conditional Analysis of Freedom’, in P. van Inwagen (ed.), *Time and Cause: Essays Presented to Richard Taylor*. Dordrecht: D. Reidel.

— 1990. *On Action*. Cambridge: Cambridge University Press.

— 2000. ‘The Epistemic Requirements for Moral Responsibility’, *Philosophical Perspectives*, 14: 267–277.

— 2006. ‘Working with Fischer and Ravizza’s Account of Moral Responsibility’, *Journal of Ethics*, 10: 229–253.

Glover, J. 1970. *Responsibility*. New York: Humanities Press.

Goertzel, B., Mossbridge, J., Monroe, E., Hanson, D., and Yu, G. 2017. 'Loving AI: Humanoid Robots as Agents of Human Consciousness Expansion', *arXiv preprint arXiv:1709.07791*.

Goya-Martinez, M. 2016. 'Chapter 8 - The Emulation of Emotions in Artificial Intelligence: Another Step into Anthropomorphism', in S.Y. Tettegah and S.U. Noble (eds.) *Emotions, Technology, and Design*. London: Elsevier Inc.

Graham, P.A. 2008. 'A Defense of Local Miracle Compatibilism', *Philosophical Studies*, 140: 65–82.

Gunkel, D.J. 2012. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. Cambridge, Mass: MIT Press.

— 2017. 'Mind the Gap: Responsible Robotics and the Problem of Responsibility', *Ethics and Information Technology*, 22(4): 307–320.

Gurney, J.K. 2015. 'Driving into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles', *Wake Forest Journal of Law and Policy*, 5 (2): 393-442.

Gurney, J.K. 2016. 'Crashing into the Unknown: An Examination of Crash-Optimization Algorithms through the Two Lanes of Ethics and Law', *Albany Law Review*, 79 (1): 183-267.

Haikonen, P.O. 2007. *Robot Brains: Circuits and Systems for Conscious Machines*. Chichester, UK: John Wiley.

Haji, I. 2002. 'Compatibilist Views of Freedom and Responsibility', in R. Kane (ed.) *The Oxford Handbook of Free Will.* New York: Oxford University Press.

Heidegger, M. 1977. *The Question Concerning Technology and Other Essays* (trans. by William Lovitt). New York: Harper and Row.

Hellström, T. 2013. 'On the Moral Responsibility of Military Robots', *Ethics and Information Technology*, 15: 99–107.

Hevelke, A. and Nida-Rümelin, J. 2014. 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis', *Science and Engineering Ethics,* 21: 619–630.

Hoefer, C. 2016. 'Causal Determinism', in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), URL = <https://plato.stanford.edu/archives/spr2016/entries/determinism-causal/>.

Holton, R. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Horgan, T. 1985. 'Compatibilism and the Consequence Argument', *Philosophical Studies*, 47(3): 339–356.

Hornby, G.S., Takamura, S., Yamamoto, T., and Fujita, M. 2005. 'Autonomous Evolution of Dynamic Gaits with Two Quadruped Robots', *IEEE Transactions on Robotics*, 21 (3): 402-410.

Huemer, M. 2000. 'Van Inwagen's Consequence Argument', *Philosophical Review*, 109: 525-44.

— 2004. 'Elusive Freedom? A Reply to Helen Beebee', *Philosophical Review*,113 (3): 411-416

Hume, D. 1975. *An Enquiry Concerning Human Understanding*, P.H. Nidditch (ed.). Oxford: Clarendon Press.

Huxley, A. 1932. *Brave New World.* London: Vintage Publishing.

Hyslop, A. 1995. *Other Minds*. Dordrecht: Kluwer Academic Publishers.

International Organization for Standardization (ISO). 2012. *ISO-Standard 8373:2012 Robots and Robotic Devices – Vocabulary.* ISO [Online]. Available at: https://www.iso.org/obp/ui/#iso:std:iso:8373:ed-2:v1:en (Accessed November 12, 2020).

van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Clarendon Press

— 1989. 'When is the Will Free?', *Philosophical Perspectives, 3*: 399–422.

— 1997. 'Fischer on Moral Responsibility', *The Philosophical Quarterly,* 47 (188): 373–381

— 1999. 'Moral Responsibility, Determinism and the Ability to do Otherwise', *The Journal of Ethics,* 3: 343-351.

— 2000. 'Free Will Remains a Mystery: The Eighth Philosophical Perspectives Lecture', *Philosophical Perspectives,* 14: 1–19.

— 'Freedom to Break the Laws', *Midwest Studies in Philosophy*, 28: 334-350.

I, Robot. 2004. Directed by Alex Proyas. [Feature film]. Los Angeles, CA: 20th Century Studios.

Jefferson, A. 2019. 'Instrumentalism about Moral Responsibility Revisited', *The Philosophical Quarterly*, 69 (276): 555-573.

Johnson, D.G. 2006. 'Computer Systems: Moral Entities but not Moral Agents', *Ethics and Information Technology*, 8: 195-204.

Kane, R. 1989. 'Two Kinds of Incompatibilism', *Philosophy and Phenomenological Research,* 50 (2): 219–254.

— 1996. *The Significance of Free Will*. Oxford: Oxford University Press.

— 2005. *A Contemporary Introduction to Free Will.* New York: Oxford University Press.

Kim, Y. and Shyam Sundar, S. 2012. 'Anthropomorphism of Computers: Is it Mindful or Mindless?', *Computers in Human Behavior*, 28 (1): 241-250.

Kittle, S. 2015. 'Abilities to do Otherwise', *Philosophical studies*, 172 (11): 3017–3035.

Kohlberg, L. 1981. *Essays on Moral Development. Vol. I.* San Francisco: Harper & Row.

— 1984. *Essays on Moral Development. Vol. II.* San Francisco: Harper & Row.

Kozuck, B. and McKenna, M. 2016. 'Free Will, Moral Responsibility and Mental Illness', in D. Moseley and G. Gala (eds.) *Philosophy and Psychiatry: Problems, Intersections and New Perspectives.* Milton Park, England: Routledge.

Kurzweil, R. 2002. 'Locked in his Chinese Room' in J.W. Richards (ed.) *Are We Spiritual Machines: Ray Kurzweil vs. the Critics of Strong AI*. Seattle: Discovery Institute.

Lamb, J.W. 1977 'On a Proof of Incompatibilism', *The Philosophical Review*, 86 (1): 20–35.

Lancaster, K. 2019. 'The Robotic Touch: Why There is No Good Reason to Prefer Human Nurses to Carebots', *Philosophy in the Contemporary World*, 25 (2): 88-109.

Levin, J. 2021. 'Functionalism', in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition). URL = <https://plato.stanford.edu/archives/win2021/entries/functionalism/>.

Levin, S., Wong, J. 2018. 'Self-driving Uber kills Arizona Woman in First Fatal Crash involving Pedestrian', *The Guardian*, 19 March. URL=

https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe (Accessed 2 September 2018)

Levy, N. 2011. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.

— 2014. *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

—, McKenna, M. 2009. 'Recent Work on Moral Responsibility', *Philosophy Compass*, 43: 96–133.

Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.

— 1981 (2003). 'Are We Free to Break the Laws', in G. Watson (ed.) *Free Will*. Oxford: Oxford University Press.

Lin, P. 2016. 'Why Ethics Matters for Autonomous Cars', M. Maurer, J. Gerdes, B. Lenz, H. Winner (eds) *Autonomous Driving.* Berlin, Heidelberg: Springer

van Loon, R.J., and Martens, M.H. 2015. 'Automated Driving and its Effect on the Safety Ecosystem: How do Compatibility Issues Affect the Transition Period?', *Procedia Manufacturing*, 3: 3280-3285.

Lovelace, A., Menabrea, L. 1842. 'Sketch of the Analytical Engine invented by Charles Babbage, Esq.', in R. Taylor (ed.), *Scientific Memoirs*. London: Richard and John E. Taylor.

Lucas, G.R. 2013. 'Engineering, Ethics, and Industry: The Moral Challenges of Lethal Autonomy' in B.J. Strawser (ed.) *Killing by Remote Control: The Ethics of an Unmanned Military.* Oxford: Oxford University Press.

Lucas, J.R. 1995. *Responsibility*. Oxford: Clarendon

Marchant, G.E. and Lindor, R.A. 2012. 'The Coming Collision between Autonomous Vehicles and the Liability System', *Santa Clara Law Review*, 52(4): 1321–1340

Mason, E. 2015. 'Moral Ignorance and Blameworthiness', *Philosophical Studies,* 172 (11): 3037–3057.

— 2019. *Ways to Be Blameworthy: Rightness, Wrongness, and Responsibility*. Oxford: Oxford University Press.

Matheson, B. 2014. 'Compatibilism and Personal Identity' *Philosophical Studies,* 170 (2): 317–334.

Matthias, A. 2004. 'The Responsibility Gap: Ascribing responsibility for the Actions of Learning Automata', *Ethics and Information Technology,* 6 (3): 175-183.

McKay, T.J., and Johnson, D. 1996, 'A Reconsideration of an Argument against Compatibilism', *Philosophical Topics,* 24 (2): 113-122.

McCulloch, G. 1990. 'Dennett's Little Grains of Salt', *Philosophical Quarterly,* 40:1–12.

McKenna, M. 1998. 'The Limits of Evil and the Role of Moral Address: A Defence of Strawsonian Compatibilism', *The Journal of Ethics,* 2: 123-142.

— 2001. 'Review of John Martin Fischer and Mark Ravizza's Responsibility & Control' *Journal of Philosophy,* 98 (2): 93–100.

— 2012. 'Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism', *The Journal of Ethics,* 16 (2): 145–74.

— 2013. 'Reasons-responsiveness, Agents and Mechanisms', in D. Shoemaker (ed.) *Oxford Studies in Agency and Responsibility Volume 1.* Oxford: Oxford University Press

— 2016. ‘Reasons-Responsive Theories of Freedom’, in K. Timpe, M. Griffith and N. Levy (eds.) *The Routledge Companion to Free Will.* London: Taylor and Francis.

—, Coates, D.J. 2021. ‘Compatibilism’, in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), URL = https://plato.stanford.edu/archives/fall2021/entries/compatibilism/

McLaren, B.M. 2011. ‘Computational Models of Ethical Reasoning’, in M. Anderson and S.L. Anderson (eds.), *Machine Ethics*. New York: Cambridge University Press.

Mele, A.R. 1995. *Autonomous Agents: from Self-Control to Autonomy*. Oxford: Oxford University Press.

— 2000. ‘Reactive Attitudes, Reactivity, and Omissions’, *Philosophy and Phenomenological Research*, 61 (2): 447–452.

— 2006. *Free Will and Luck.* New York: Oxford University Press.

— 2008. ‘Manipulation, Compatibilism, and Moral Responsibility’, *The Journal of Ethics*, 12: 263-286.

— 2010. ‘Moral Responsibility for Actions: Epistemic and Freedom Conditions’, *Philosophical Explorations*, 13(2): 101-111.

Moore, G.E. 1912. *Ethics*. London: Williams and Norgate.

Moran, R. 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton, NJ: Princeton University Press.

Nass, C. and Moon, Y. 2000. ’Machines and Mindlessness: Social Responses to Computers’, *Journal of Social Issues,* 56 (1): 81-103.

Nelkin, D. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

Newell, A. 1982. *Intellectual Issues in the History of Artificial Intelligence.* Pittsburgh: Carnegie-Mellon University.

Noorman, M. 2018. ‘Computing and Moral Responsibility’, in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Spring 2018 Edition)*, URL = <https://plato.stanford.edu/archives/spr2018/entries/computing-responsibility/>.

Nozick, R. 1981. *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Nyholm, S. 2018. ‘The Ethics of Crashes with Self-Driving Cars: A Roadmap, II’, *Philosophy Compass*, 13: e12506

Pendergraft, G. 2011. ‘The Explanatory Power of Local Miracle Compatibilism’, *Philosophical Studies*, 156: 249-266.

Pereboom, D. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

— 2013. ’Free Will’ in R. Crisp (ed.) *The Oxford Handbook of the History of Ethics*. Oxford: Oxford University Press.

Perry, J. 2010. ‘Wretched Subterfuge: A Defense of the Compatibilism of Freedom and Natural Causation’, *Proceedings and Addresses of the American Philosophical Association,* 84 (2): 93-113

Personal Robots Group. 2008. *Official MDS Robot Video – First Test of Expressive Ability.* April 15. Available at: https://www.youtube.com/watch?v=aQS2zxmrrrA (Accessed: 27th of April 2019).

Piaget, J. 1965. *The Moral Judgment of the Child.* New York: Free Press.

Portal. 2007. PC [Game]. Washington: Valve Corporation.

Pöyhönen, S. 2014. 'Intentional Concepts in Cognitive Neuroscience', *Philosophical Explorations*, 17: 93–109.

Purves, D., Jenkins, R., and Strawser, B.J. 2015. 'Autonomous Machines, Moral Judgement, and Acting for the Right Reasons', *Ethical Theory and Moral Practise*, 18 (4): 851-872.

Putnam, H. 1965. 'Brains and Behaviour' in R.J. Butler (ed.) *Analytical Philosophy, Second Series*. Oxford: Blackwell.

— 1967 (1975). 'The Nature of Mental States' in his *Mind, Language and Reality.* Cambridge: Cambridge University Press.

Ramirez, E. 2015. 'Receptivity, Reactivity and the Successful Psychopath', *Philosophical Explorations*, 18 (3): 330-343.

Ravizza, M. 1993. 'Introduction' in J.M. Fischer and M. Ravizza (eds.) *Perspectives on Moral Responsibility.* Ithaca, NY: Cornell University Press.

Rey, G. 1994. 'Dennett's Unrealistic Psychology', *Philosophical Topics*, 22: 259–289.

— 1997. *Contemporary Philosophy of Mind*. Cambridge, MA: Blackwell.

Rigby, C. 2019. 'Starship Technologies' Robots Make Their 50,000th UK Delivery', *Internet Retailing*, 11th of April. Available at: https://internetretailing.net/delivery/delivery/starship-technologies-robots-make-their-50000th-uk-delivery--19446 (Accessed January 2, 2020).

Robb, D. 2020. 'Moral Responsibility and the Principle of Alternative Possibilities', in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), URL = https://plato.stanford.edu/archives/fall2020/entries/alternative-possibilities/

Robichaud, P. and Wieland, J.W. (eds.) 2017. *Responsibility: The Epistemic Condition*. Oxford: Oxford University Press

Roff, H. 2013. 'Killing in War: Responsibility, Liability, and Lethal Autonomous Robots', in A. Henschke, N. Evans, F. Allhoff (eds.) *Routledge Handbook for Ethics and War: Just War Theory in the 21st Century*. New York: Routledge.

Rosenthal-von der Pütten et al. 2013. 'An Experimental Study on Emotional Reactions Towards a Robot', *International Journal of Social Robotics*, 5: 17-34.

Rousseau, J. 1979. *Emile or On Education*. New York: Basic Books.

Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson.

Salmeron J.L. 2015. 'Simulating Synthetic Emotions with Fuzzy Grey Cognitive Maps', in Sinčák P., Hartono P., Virčíková M., Vaščák J., Jakša R. (eds.) *Emergent Trends in Robotics and Intelligent Systems. Advances in Intelligent Systems and Computing.* Switzerland: Springer, Cham.

Sartorio, C. 2015. ‘Sensitivity to Reasons and Actual Sequences’, in D. Shoemaker (ed.) *Oxford Studies in Agency and Responsibility: Volume 3.* Oxford: Oxford University Press.

Scarantino, A., and de Sousa, R. ‘Emotion’, in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2021). URL = <https://plato.stanford.edu/archives/sum2021/entries/emotion/>.

Schoettle, B. and Sivak, M. 2015. *A Preliminary Analysis of Real-World Crashes Involving Self-Driving Vehicles (no. UMTRI-2015-34).* Ann Arbor: The University of Michigan Transportation Research Institute.

Sharkey, A. and Sharkey, N. 2012. ‘Granny and the Robots: Ethical Issues in Robot Care for the Elderly’, *Ethics and Information Technology*, 14: 27-40.

Silver, D., Hubert, T., et al. 2018. ‘A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go through Self-Play’, *Science,* 362 (6419): 1140-1144.

—, Schrittwieser, J., et al. 2017. ‘Mastering the Game of Go without Human Knowledge’, *Nature,* 550: 354–359.

Skinner, B.F. 1974. *About Behaviorism*. New York: Vintage.

Slors, M. V. P. 2007. ‘Intentional Systems Theory, Mental Causation and Empathic Resonance’, *Erkenntnis,* 67: 321–336.

Sparrow, R. 2007. ‘Killer Robots’, *Journal of Applied Philosophy,* 24(1): 62–77.

Shoemaker, S. 1990. ‘Review’, *The Journal of Philosophy,* 87 (4): 212-216.

Stahl, B. C. 2004. ‘Information, Ethics and Computers: The Problem of Autonomous Moral Agents’, *Minds and Machines*, 14: 67–83.

Stout, N. 2016. ‘Reasons-Responsiveness and Moral Responsibility: The Case of Autism’, *Journal of Ethics*, 20: 401–418.

Strawson, G. 1986. *Freedom and Belief*. Oxford: Clarendon Press.

— 1994. ‘The Impossibility of Ultimate Moral Responsibility’, *Philosophical Studies*, 75: 5-24.

Strawson, P. 1962 (2003). ’Freedom and Resentment’ in G. Watson (ed.) *Free Will*. Oxford: Oxford University Press.

Sullins, J. P. 2006. ‘When is a Robot a Moral Agent?’, *International Review of Information Ethics*, 6 (12), 23–30.

Taylor, R. 1974. *Metaphysics*. Englewood Cliffs, N.J.: Prentice Hall.

Todd, P. 2016. ‘Strawson, Moral Responsibility, and the “Order of Explanation”: An Intervention’, *Ethics*, 127: 208-240.

U.S. Department of Defense. 2011. ‘Unmanned Systems Integrated Roadmap FY 2011-2036’. Online. URL= http://www.defenseinnovationmarketplace.mil/resources/UnmannedSystemsIntegratedRoadmapFY2011.pdf (accessed 11 February, 2017).

Vargas, M. 2006. ‘On the Importance of History for Responsible Agency’, *Philosophical Studies*, 127(3): 351-382.

Vargas, M. Forthcoming. ‘Instrumentalist Theories of Moral Responsibility’, in D.K. Nelkin and D. Pereboom (eds.) *The Oxford Handbook of Moral Responsibility*. Oxford: Oxford University Press.

Velleman, D.J. 2002. ‘Identification and Identity’, in S. Buss and L. Overton (eds.) *Contours of Agency: Essays on Themes from Harry Frankfurt*. Cambridge, Mass.: MIT Press.

Wallace, R.J. 1996. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

— 1997. ‘Review of John Martin Fischer’s The Metaphysics of Free Will’, *Journal of Philosophy*, 94: 156–159.

Wallach, W. and Allen, C. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

—, Allen, C. and Smit, I. 2008. ‘Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties’, *AI & Soc*, 22: 565–582.

Watson, G. 1975. ‘Free agency’, *Journal of Philosophy,* 72: 205–220.

— 1987. ‘Responsibility and the Limits of Evil: Variations on a Strawsonian Theme’, in Schoeman, F.D. (ed.) 1987. *Responsibility, Character and the Emotions: New Essays in Moral Psychology*. Cambridge: Cambridge University Press.

— 1999. ‘Soft Libertarianism, Hard Compatibilism’, *Journal of Ethics*, 3 (4): 351-365.

— 2001. ‘Reason and Responsibility’, *Ethics*, 111: 374–394.

Watson, S., Duecker, D.A. and Groves, K. 2020. 'Localisation of Unmanned Underwater Vehicles (UUVs) in Complex and Confined Environments: A Review', *Sensors*, 20 (21):1–35.

Weiss, L.G. 2011. 'Autonomous Robots in the Fog of War', *IEEE Spectrum*, 48(8): 30–57.

Weizenbaum, J. 1976. *Computer Power and Human Reason*. San Francisco, CA: Freeman and Co.

Wiggins, D. 1973. 'Towards a Reasonable Libertarianism', in T. Honderich (ed.) *Essays on Freedom of Action*. Boston: Routledge and Kegan Paul.

Wolf, S. 1987 (2003). 'Sanity and the Metaphysics of Responsibility' in G. Watson (ed.) *Free Will*. Oxford: Oxford University Press.

Wolf, S. 1990. *Freedom Within Reason*. Oxford: Oxford University Press.

— 2001. 'The Moral of Moral Luck', *Philosophic Exchange*, 31: 4–19.

Wortham, R.H. (2020). *Transparency for Robots and Autonomous Systems: Fundamentals, Technologies and Applications*. London: The Institution of Engineering and Technology.

Yaffe, G. 2003. 'Indoctrination, Coercion and Freedom of Will', *Philosophy and Phenomenological Research*, 67 (2): 335-356.

Yang, Q., Gao, Y., and Li, Y. 2016. 'Suppose Future Traffic Accidents Based on Development of Self-Driving Vehicles', in S. Long and B.S. Dhillon (eds.) *Man-Machine-Environment System Engineering: Lecture Notes in Electrical Engineering.* New York: Springer.

Zimmerman, D. 2003. 'That Was Then, This Is Now: Personal History vs. Psychological Structure in Compatibilist Theories of Autonomous Agency', *Noûs*, 37 (4): 638-671.

Zimmerman, M.J. 2010. ‘Responsibility, Reaction, and Value’, *Journal of Ethics*, 14: 103–115.