Ethics and Information Technology (2025) 27:12
https://doi.org/10.1007/s10676-025-09820-x

ORIGINAL PAPER

Designing responsible agents

Zacharus Gudmunsen[1]

Published online: 4 February 2025
© The Author(s) 2025

Abstract
Raul Hakli & Pekka Mäkelä (2016, 2019) make a popular assumption in machine ethics explicit by arguing that artificial agents cannot be responsible because they are designed. Designed agents, they think, are analogous to manipulated humans and therefore not meaningfully in control of their actions. Contrary to this, I argue that under all mainstream theories of responsibility, designed agents can be responsible. To do so, I identify the closest parallel discussion in the literature on responsibility and free will, which concerns 'design cases'. Design cases are theoretical examples of agents that appear to lack responsibility because they were designed, philosophers use these cases to explore the relationship between design and responsibility. This paper presents several replies to design cases from the responsibility literature and uses those replies to situate the corresponding positions on the design and responsibility of artificial agents in machine ethics. I argue that each reply can support the design of responsible agents. However, each reply also entails different levels of severity in the constraints for the design of responsible agents. I offer a brief discussion of the nature of those constraints, highlighting the challenges respective to each reply. I conclude that designing responsible agents is possible, with the caveat that the difficulty of doing so will vary according to one's favoured reply to design cases.

Keywords Artificial moral agency · Control · Design · Machine ethics · Manipulation · Responsibility

Introduction

Machine ethics is a growing field that centres on answering questions about how and whether artificial agents can possess moral qualities or abilities. Here, I use 'agent' in a loose (roughly interpretivist) sense, without requiring that agents have intentional or conscious internal states. Artificial agents are demarcated from 'natural' agents partly by being products of a *design* process. Design processes may be collaborative or individual, simple or complex, and result in a wide variety of agents and objects. I call any agent that is the product of a design process a 'designed agent'. All artificial agents today seem to be designed agents, and philosophical worries that designed agents are unable to have all the moral qualities that humans do are, as I will discuss, an important topic in machine ethics.

This brings me to the main question of the paper: how should our understanding of *design* inform machine ethics? The obvious area to focus on here is *responsibility*, *freedom*, and *self-control.* After all, if anything is going to go awry for designed agents, it will be that they have less control, freedom or autonomy (and therefore less responsibility) compared to their undesigned counterparts.

Machine ethicists are split on the issue. Some assume that there is a fundamental tension between being designed and being responsible. But others argue that any tension between the concepts must be illusory, because humans are analogously 'designed' and remain responsible. Here, I aim to situate these positions within the ongoing debate on responsibility and manipulation.

The relationship between design and responsibility is central to discussions concerning 'design cases', a subset of 'manipulation cases' in the literature on responsibility. 'Design cases' are theoretical examples intended to probe the relationship between design and responsibility - they pose hypothetical cases in which an agent is designed and appears to lack responsibility. Most philosophers of responsibility take a 'hard-line' reply to design cases – they claim that the only restriction on the type of design that can produce responsible agents is that the resulting agent satisfies all the normal conditions for responsibility. But a growing group

✉ Zacharus Gudmunsen
sgudmunsen@ku.edu.tr

[1] Koç University, Istanbul, Turkey

of philosophers have been supporting 'soft-line' replies. On these views, designing a responsible agent requires a further condition: the design process itself must not be a responsibility-undermining causal source of the designed agent's actions. While some may argue that designing an agent entails that the designer is just such a responsibility-undermining causal source, I argue against this. I claim that if certain conditions are met, designing an agent may not be a responsibility-undermining causal source of the designed agent's actions.

The structure of the paper is as follows. It first introduces the opposed positions on design and responsibility in machine ethics. Then it situates these positions within the literature on responsibility by comparing them with 'design cases'. Following this, it offers several replies to design cases from the literature, again relating these back to the positions in machine ethics Finally, it discusses each reply's resulting design constraints for producing a responsible agent. This leads to the suggestion that responsible agents can be designed, but that a designer needs to carefully satisfy certain challenging conditions to do so.

Two opposing views on design and responsibility in machine ethics

Raul Hakli & Pekka Mäkelä (2016, 2019) suggest[1] that no designed agent can be responsible in the following passage:

> [E]ven though robots could be programmed and engineered to have all the capacities required of moral agents, they would still not be moral agents, because such programming and engineering is closer to […] manipulation […] It is precisely the fact that the responsibility-relevant property X was engineered that undermines the responsibility attribution to the agent: Robots cannot be morally responsible because they are designed and programmed by other agents to have the "character" they have. (Hakli & Mäkelä, 2016)

For simplicity, let us assume that the claim here is that designed agents cannot be responsible *in general*. That is, designed agents may be responsible for some actions in limited circumstances, but they are not responsible for most of their actions or for actions carried out in routine circumstances.

As Hakli and Mäkelä discuss, it is generally accepted that manipulation violates a condition of responsibility. Martin Fischer and Mark Ravizza (1998) call the violated condition 'guidance control'. More generally, responsibility is typically broken down into 'control' and 'epistemic' conditions. Responsibility for an action requires an agent to be *in control* of its performance and *aware* of its consequences. The focus here is on the 'control' side of things: an agent that has been manipulated fails to be relevantly in control.

Whether design is like manipulation depends on how 'design' is defined. Here, I understand 'design' to be the intentional modification or creation of a physical material that confers a sufficiently predictively effective function upon that material. Various aspects of this definition are defended in the literature on design (Borgo et al., 2014; McLaughlin, 2000; Preston, 2009; Vermaas et al., 2013; Vermaas & Houkes, 2006). I will not defend it in detail here – different definitions of design may change the set of designed and responsible agents, but I focus on agents designed in this sense.

One issue to highlight in this definition is that the level of predictive efficacy required is vague. For my purposes, designing something involves conferring a function that adequately explains *most* of a system's behaviour or usage. Creating an object that behaves in accordance with a unintended or unanticipated function most of the time is not designing that object. Though there is room for variance in the specificity of the designed function, which may be quite general (such as designing a material to be strong). Another caveat here is that I limit the scope of 'design' to 'initial' design rather than 'revisionary' or 'derivative' design[2] – i.e., 'design' refers to the *first instance* of design, but not occasions when you redesign an already designed object (See Millikan, 1999).

Hakli & Mäkelä's claim is that design and manipulation are closely analogous. They suggest that design violates the same condition of responsibility that manipulation does. Concluding that designed agents cannot be responsible – a conclusion that maintains wide support within machine ethics, particularly with relation to so-called 'responsibility gaps' (Cavalcante Siebert et al., 2023; Danaher, 2016, 2022; Köhler et al., 2018; Matthias, 2004; Oimann, 2023; Santoni de Sio & van den Hoven, 2018; Vallor & Vierkant, 2024; Veluwenkamp, 2022).

However, this is met by a powerful counterargument (Danaher, 2020; Himma, 2009, p. 23; Ioan & Howard, 2017; Musiał, 2017; Petersen, 2011; Powers, 2013, p. 235; Véliz, 2021). Behdadi and Munthe summarise the claim as the following.

[1] Several others can be interpreted as sympathetic to this claim. In military robotics, some think that artificial systems cannot be autonomous (Hellström, 2013; Schulzke, 2013) and design seems to play *some* role in why. Similarly, others in machine ethics suggest that artificial agents cannot be autonomous because they are programmed by designers (Bringsjord, 2008; Grodzinsky et al., 2008; Johnson, 2006; Torrance, 2008). Chomanski (2019) and Musiał (2017, 2022) claim that designing agents is *manipulative* and reappear later.

[2] Without this caveat the lines between 'manipulation' and 'design' become blurred – redesigning an agent is no different to straightforward manipulation and uncontroversially undermines responsibility.

> [A] condition that excludes machines based on being designed will fail to distinguish between human and artificial entities with regard to moral agency, thus undermining the notion of human beings as paradigmatic moral agents. (Behdadi and Munthe, 2020)

According to this line of reasoning, design may not violate any condition of responsibility because any responsibility-undermining quality of design also applies to humans.

I will return to that argument later. The point for now is that there are two opposing positions on designed agents' responsibility: Hakli & Mäkelä represent the position that designed agents cannot be responsible while Behdadi & Munthe present the position that design is not in tension with responsibility.

Design cases

I will now present a few examples intended to elicit the intuition that design is relevant to responsibility. These cases arose within the debate about the conditions of responsibility. They are scenarios in which agents are designed in a way that appears to undermine responsibility[3].

The first case is Alfred Mele's zygote case.

> Diana creates a zygote Z in Mary. She combines Z' s atoms as she does because she wants a certain event E to occur thirty years later. From her knowledge of the state of the universe just prior to her creating Z and the laws of nature of her deterministic universe, she deduces that a zygote with precisely Z' s constitution located in Mary will develop into an ideally self-controlled agent who, in thirty years, will judge, on the basis of rational deliberation, that it is best to A and will A on the basis of that judgment, thereby bringing about E. (Mele, 2008, p. 188)

Zygote Z becomes a person named Ernie who brings about event E. Mele claims that Ernie is 'neither free nor responsible'. But Mele's initial formulation of the case is not design as I have defined it. Diana does not confer a predictively effective function to Ernie. The function Diana confers is to perform a single action – it does not predict any of his other actions. However, Mele, anticipating something like this, suggests modifying the case such that Diana creates Ernie with the intention that he perform *every* action. Let's move forward with this modified version.

Diana, then, designs Ernie. He is contrasted with 'Bernie', a version of Ernie who performs the same actions without Diana intervening. It must be the case that Bernie is responsible, because he is no different to an ordinary human[4]. In contrast, it is intuitive to many that Ernie is not responsible at all. If so, the best explanation seems to be that Ernie is not responsible because he was designed, as that is the only difference between him and Bernie.

The second design case comes from Derk Pereboom's four case argument. In the first case of the four-case argument, scientists manipulate a person called Plum (call him Plum1) to kill a person called White. This manipulation occurs around the same time that Plum1 actually commits the deed, so it is a case of straightforward manipulation, not design. But the second case of the four-case argument resembles design:

> Case 2: Plum[2] is just like an ordinary human being, except that a team of neuroscientists programmed him at the beginning of his life so that his reasoning is often but not always egoistic […] and at times strongly so, with the intended consequence that in his current circumstances he is causally determined to engage in the egoistic reasons responsive process of deliberation and to have the set of first and second-order desires that result in his decision to kill White. Plum[2] has the general ability to regulate his actions by moral reasons, but in his circumstances, due to the strongly egoistic nature of his deliberative reasoning, he is causally determined to make his decision to kill. Yet he does not decide as he does because of an irresistible desire. (Pereboom, 2001, pp. 113–114)

In this second case the scientists intervened at the beginning of Plum's (call him Plum2) life and influenced his dispositions more generally. Plum2's programmed function effectively predicts much of his actions because his reasoning is *often* egoistic.[5] So the neuroscientists appear to have designed Plum2.

According to Pereboom, Plum2 is "not morally responsible because his action is determined by the neuroscientists' programming, which is beyond his control." (Pereboom, 2001, p. 114). Pereboom thinks that Plum1 is not responsible and that there is no relevant difference between Plum1 and Plum2. He writes: "Furthermore, it would seem unprincipled to claim that […] Plum[2] is morally responsible

[3] The first two cases about to be presented have been subject to a large amount of discussion. I discuss some constructive replies to them in the following sections, but I do not take on the task of defending the validity of design cases in general (For discussion see, for example, Kearns, 2012; King, 2013; Matheson, 2016; Mickelson, 2015).

[4] I assume here that Strawsonian style arguments against human responsibility are to be rejected and that ordinary humans are responsible in general (See Latham & Tierney, 2022).

[5] It might still be an open question whether this egoistic disposition effectively predicts enough of Plum's actions such that he is *designed*, but let's assume that it does.

because the length of time between the programming and his decision is now great enough. Whether the programming occurs a few seconds before or forty years prior to the action seems irrelevant…" (Pereboom, 2001, p. 114).

Suppose that Pereboom is right that neither Plum1 nor Plum2 is responsible for *killing White*. Even in this case, their wider capacities for responsibility seem to vary. Plum2 is also not responsible for many other actions, since he performs them due to his generally egoistic disposition. The difference between Plum1 and Plum2 is not merely timing here. It is that Plum2 has been implanted with a general behavioural disposition while Plum1 has not. Plum1 is responsible for all his actions other than killing White, but Plum2 is, seemingly, not responsible in general. If so, the reason he lacks responsibility is because he was designed.

The final design case comes from literature about whether we ought to design 'artificial persons' to serve humans. Walker (2006) argues that designing artificial persons to serve humans would undermine their autonomy. But his argument is that artificial servants lack autonomy because they are designed *to be servants* rather than because they are *designed*. Steve Petersen (2011), Bartek Chomanski (2019) and Musiał (2022) suggest that *designing* artificial servants is manipulative. Chomanski explains:

> To see why programming AIs to be servants is manipulative, consider the following points. AI servants are artificial people programmed to find the deepest fulfillment in attending to human needs. By programming AIs with overwhelming desires to be servants, their programmers are set on making it close to impossibly psychologically difficult for the AIs to act in any way other than to pursue the life-plan that they are given: even a determined AI servant would likely not pick a different career path. This might be so even if they are programmed with the capacity to reflect on their own desires. Upon being told to build an AI servant, the programmers cannot help but suppose that the servant's decision about what life plan to pursue are for the programmers to make. […] The programmers position themselves so as to be able to steer the choices of AI servants with a high degree of efficacy […] The AIs' number of options, when it comes to career choice, is reduced to basically one. (Chomanski, 2019)

This seems to be a design case. At least, if Ernie and Plum2 lack responsibility, then there seem no relevant differences in the case of artificial servants. If anything, there are even fewer grounds for responsibility – Ernie and Plum2 are plausibly identical to some possible non-designed agent, whereas there are no natural agents identical to artificial servants.

These cases demonstrate that design can undermine responsibility on occasion. But more needs to be done to show that *design* is the problem, rather than some contingent aspect of the design process that happens to be at play in these sorts of cases. Nonetheless, design clearly has some relevance to responsibility. The concern is the exact nature of the relationship – what types of design are responsibility-undermining and why? I want to use the theoretical responses to these cases as a guide to understanding the relationship between design and responsibility. The conditions for responsibility can be explored relative to those responses, rather than fleshing out cases further. But design cases serve to contextualise the discussion – we can ask ourselves how effective the replies are at explaining the (lack of) responsibility in design cases.

Hard-line replies to design cases

Drawing on Michael Mckenna's (2008) categorisation of two types of compatibilist response to manipulation cases, I will likewise split the responses to design cases. The 'hard-line' reply claims that designed agents can be responsible because they may not be relevantly different to responsible agents; the 'soft-line' reply claims that design may well constitute a relevant difference between designed and responsible agents.

But I want to first acknowledge incompatibilist theories of responsibility. Incompatibilists about free-will might claim that to be responsible one must be free, and that, therefore, Plum2 and Ernie are responsible only if they have free-will. Pending a full explanation of what it means to be free, there may be a satisfying explanation here. Although it has been suggested that incompatibilists ought to retain some interest in the hard-line/soft-line replies (McKenna, 2016, pp. 87–88). To keep my discussion a reasonable length I will put incompatibilism to one side and proceed assuming that compatibilism is true.

The hard-line reply is popular in straightforward manipulation cases, but it is even more popular in design cases. Harry Frankfurt supports a hard-line reply to both manipulation and design cases in the oft-cited passage:

> We are inevitably fashioned and sustained, after all, by circumstances over which we have no control. The causes to which we are subject may also change us radically, without thereby bringing it about that we are not morally responsible agents. … We are the sorts of persons we are; and it is what we are, rather than the history of our development, that counts. (Frankfurt, 2002, p. 28)

> There is no paradox in the supposition that a [manipulator] might create a morally free agent. (Frankfurt, 1988, p. 54)

Watson claims that we must insist "that free agents might indeed be the products of manipulation by designers". He likewise defends a hard-line reply to both design and manipulation cases.

> What seems in the abstract like a responsibility-undermining history might seem so only because it abstracts from the constitutive proper ties of what is supposed to emerge (by design or not) from that history. If we fill out these histories, according to compatibilism, some will be responsibility-undermining, some not. If not, that will be because the design fails to realize some of the [conditions of responsibility], not because it is deterministic. (Watson, 1999, 365).

But some treat design and straightforward manipulation cases differently. Fischer ordinarily supports a soft-line reply to manipulation, but shifts to a hard-line reply for design cases.

> Elsewhere, I have argued that the intuition that [Plum1] is not in fact morally responsible in Case 1 issues from a view that the mechanism that leads to his behavior in that case is not his own (Fischer, 2014). But this view cannot be applied straightforwardly to Case 2; although the neuroscientists preprogram [Plum2] to have the mechanisms of practical reasoning that he has, and to act on them in certain specific circumstances, this does not imply that they are not *his* mechanisms. (Fischer, 2017, pp. 235–236)

Fischer takes a soft-line reply to Plum1 – Plum1 is not responsible because manipulation violates an ownership condition for responsibility; and a hard-line response to Plum2 – Plum2 meets the ownership condition because the intervention is prior to birth, making him responsible (and he holds the same position for the zygote case; Fischer, 2011, 2016). In taking a hard-line reply to design cases, Fischer agrees with Frankfurt and Watson about design. They agree that Ernie and Plum2 *are* responsible. Mele, who has argued for a soft-line reply to straightforward manipulation cases, agrees that compatibilists should accept that "when it comes to moral responsibility, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe." (Mele, 2013, p. 180) and thus that "Ernie should strike them as free and morally responsible (in light of his properties as an agent)." (Mele, 2008, p. 193).

Note that Mele and Fischer have a particular reason for preferring a hard-line reply to design cases and a soft-line reply to manipulation. To them, manipulation constitutes a relevant difference because the manipulated agent's original beliefs and values are *overridden*. In design cases there is no overriding and thus, they think, no relevant difference between designed agents and non-designed, naturally occurring agents. Both result in mechanisms that are 'one's own'.

It is this 'hard-line' position that represents Behdadi & Munthe's side of the debate in machine ethics. Those who support the 'standard objection' that 'excluding designed agents from responsibility will also exclude humans' are offering the hard-line reply just described.

Hakli and Mäkelä defend against this reply, but their defence appears mistargeted. Defending their position that design undermines responsibility requires a soft-line reply. But Hakli and Mäkelä attempt to settle things by appealing to compatibilism.

> Our claim is not that being externally determined undermines autonomy, and mutatis mutandis, moral agency. We are not just giving a rerun to the reasoning that robots are programmed, hence not free, hence not autonomous which has already been discussed in the literature (see, e.g., Bringsjord [2008]) and to which there is a standard objection too (see, e.g., Sullins [2006]): the "paradigmatic moral agents," that is, adult human beings are bound to face the same problem as robots. This is because human beings are what they are due to their genetic programming, upbringing, and contingent influences from their environment. [This 'standard objection' is the hard-line reply]
>
> However, in our view, the possibility that human beings' characters are "determined" by external factors and influences is not the decisive factor in deciding whether an agent is autonomous and responsible. At the background of history-sensitive externalism there is a compatibilist idea according to which autonomy is not necessarily ruled out by determination. Indeed the main point is not about determination per se, rather the relevant issue is the kind of determination: An agent can be autonomous even in a deterministic universe as long as their choices result from their own authentic goals and values. (Hakli & Mäkelä, 2019, 269)

Hakli and Mäkelä here seem to respond to incompatibilists. They claim that, since compatibilism is true, they may adopt the soft-line reply to design cases. But the hard-line reply to design cases need not assume incompatibilism. Compatibilists like Frankfurt and Fischer both leverage the claim about there being no difference between responsible agents

and some designed agents. They adopt a hard-line reply to design cases while accepting that 'an agent can be autonomous even in a deterministic universe'.

The relevant issue is whether a designed agent can make choices that 'result from their own authentic goals and values'. Fischer and Frankfurt think that they can, while Hakli & Mäkelä think that they cannot. The 'standard objection' in machine ethics summarised by Behdadi and Munthe is best construed as a compatibilist position. It supports the hard-line reply to design cases by denying a responsibility relevant difference between designed and naturally occurring agents. Its backers do not tend to argue that designed agents lack incompatibilist free-will (though Johnson (2006) is perhaps an exception).

So, one option on the table for guiding the design a responsible agent is the hard-line reply to design cases. On this view, designed agents can be responsible so long as they bear no relevant difference to some other responsible agent. I'll discuss the design constraints that follow from adopting this reply in the final section.

The hard-line reply is popular, but it remains counter-intuitive. Imagine a genetically modified group of humans who are designed to meet the conditions of responsibility but are imbued with a disposition to be vicious, hierarchical, violent and power-driven. Surely, we must think, as in the artificial servants, Plum2 and Ernie cases, they are not *as* responsible as undesigned humans. They lack responsibility, it seems, because it's not up to *them* to be violent and aggressive. But according to the hard-line reply they *are* just as responsible because their aggressive dispositions were implanted at birth rather than overridden. They are just as responsible as a human who developed the same characteristics naturally.

Those who accept the hard-line reply to design cases do so at the cost of accepting that responsibility doesn't always follow our intuitions – an agent might intuitively seem to lack responsibility when they are in fact responsible. This is also difficult to explain – returning to Pereboom's point about timing: why is overriding a *future* disposition compatible with responsibility while overriding an *existing* disposition is not? Those who don't want to accept those costs can turn to the 'soft-line' reply.[6]

Soft-line replies to design cases

The general theme of the soft-line reply is that responsibility-undermining interventions have distinctive qualities independent of whether they *override* a disposition. One theory that takes this approach claims that responsibility-undermining interventions are distinct because they are the 'causal sources' of actions. Another claims that they are distinct because they are the product of a manipulating agent's *effective intentions*.

Oisín Deery & Eddy Nahmias (2017) offer an account of causal sourcehood[7] that is intended to facilitate a soft-line reply to design cases. On this account, "[A]n agent can be the causal source of her actions, since often no variable beyond the agent's control will have a stronger causal–explanatory relationship with her actions than relevant variables within her control. On the other hand, the causal source of a manipulated agent's actions lies beyond the agent's control in the intentions of the manipulator." (Deery & Nahmias, 2017).

Deery & Nahmias claim that responsible agents must be the causal source of their actions. Being a causal source involves two strong relations: first, the presence and absence of the cause strongly correlates with the presence and absence of the action; second, there is no alternative causal source that has a stronger causal relationship with the action. Actions may have more than one causal source if they have two equally powerful causes. But, according to Deery & Nahmias, most actions have only one causal source.

Under this model, an agent is the causal source of their own action if their internal mechanisms have the strongest causal relationship with the action. For example, you are the causal source of your travelling home if your decision-making process has the strongest causal relationship with that outcome. That is, if you decided *not* to travel home you would, in most circumstances, not travel home, and if you decide *to* travel home you would, in most circumstances, travel home. Furthermore, there is no more powerful cause of your travelling home – for example, there is no kidnapper determined to take you home regardless of your decision. There may be other contributing causes: you might travel home because you feel tired or because it is 5pm, but so long as these variables are less causally efficacious than your decision[8], they are not the *causal source* of your action.

Deery and Nahmias (2017) state that "[i]t is relevant to an agent's free will and moral responsibility for performing an action, A, whether A has its causal source in the agent". On this account, designers are likely the causal source of most of designed agents' actions. In contrast, responsible agents are typically the causal source of their own actions. However, Deery & Nahmias' account sometimes fails to ascribe responsibility when it intuitively should. Hannah Tierney

[6] Arguments about the merits of hard and soft-line replies extend much further than discussed here, but my intention is not to settle the issue, but rather to summarise the views to make their correspondence with the positions in machine ethics clear.

[7] Drawn from Woodward (2005, 2015).

[8] That is, you may not return home despite it being past 5pm and/or you being tired, whereas you rarely (or never) return home without deciding to do so.

& David Glick (2020) make this complaint convincingly[9]. Roughly, the idea is that sometimes we are still responsible for actions even when we are not their causal source, Tierney & Glick use the example of a pothole in the road being the causal source of a car crash. Even in this case, they say, there are cases in which we still hold the driver responsible for the crash.

One way of accommodating this is to treat agents and non-agents differently. Perhaps agents are more responsibility-undermining than non-agents. Agents' actions may have a greater weight so that even in cases like the pothole example the driver remains responsible. Deery and Nahmias (2017), along with Marius Usher (2018), think that the right model of causation will have it that agents are more often causal sources than non-agents because agents' ability to plan and adapt to circumstances means that they cut a straighter modal line toward particular outcomes than non-agents. That is, agents are generally stronger causal forces than non-agents – using this kind of approach would suggest that the pothole was not the causal source of the crash, realigning with our typical attribution of responsibility.

Even if privileging agents as causal sources is not entailed by the right causal model, some other soft-line replies focus on agents alone as responsibility-undermining influences. Liu (2022) calls these 'another-agent' views; De Marco (2023) calls them 'manipulator' views. They argue that the influence of other *agents* is key to the responsibility-undermining force of design cases.

This is not a new idea; it first appears in A. J. Ayer's (1963) defence of compatibilism; and reoccurs in Dennett (1984). But a more contemporary example, Waller (2014), argues that Ernie lacks responsibility because his actions are caused by Diana's 'effective intention'. As only agents have intentions, this prioritises agents as responsibility-undermining causes.

Waller expresses the relationship between effective intentions and responsibility like so:

> S is less deserving of blame or praise for A-ing than she would be otherwise if (1) another agent G effectively intends that S A-s, (2) G brings it about that S A-s via intervention, and (3) S did not intentionally bring it about that G intends that S A-s. (Waller, 2014).

On Deery & Nahmias' model, the relevant sense in which G brings it about that S A-s via intervention is when G's action is the causal source of S A-ing. But the salient difference here is that Waller focuses on *agents* specifically – G must also 'effectively intend' that S A-s.

The focus on intention can be problematic – what does it take for an agent to 'intend' something? A full explanation of mental states might imply that 'intentions' and 'plans' reduce into behavioural dispositions, that intentions are vague and subject to interpretation, or are causally inert – any of which may spell trouble for this condition. Similarly, accidental actions may be clearly causal sources and intuitively undermine others' responsibility without being effective intentions, meaning that there is likely to be a good case for supplementing 'another-agent' accounts with at least some appeal to causal-sourcehood.

But, while there are outstanding questions for every reply to design cases, another-agent and causal-sourcehood accounts can provide an account in which design is in tension with responsibility. Designers are other agents and are often the causal sources of designed agents' actions.

Constraints for designing responsible agents

So far, I've raised two kinds of reply to design cases – hard-line replies and soft-line replies. Hard-line replies appear to reflect the position that design is not in tension with responsibility, while soft-line replies appear to reflect the best interpretation of Hakli & Mäkelä's claim that designed agents cannot be responsible. Here, I will argue here that both replies can endorse the possibility of designing responsible agents. This section takes them in turn, first defending the idea that designed agents can be responsible and then identifying the constraints on their design that each reply imposes.

On the hard-line reply to design cases, a designed agent can be responsible so long as they are not otherwise relevantly different to a responsible agent. On a Frankfurt/Watsonian account, this involves having the right agential structure – the arrangement of mental states that corresponds to authentically endorsing one's actions. Any agent that achieves this, regardless of its origins, is held to be responsible.

The challenge for would-be designers of responsible agents is achieving this. The design constraints implied by the hard-line reply appear to be permissive – there are no restrictions on the kind of design that can produce a responsible agent. But some accounts are more restrictive than they appear. The Frankfurt/Watson hard-line reply appeals to non-designed agents as prototypical exemplars of responsible agents, and then argues that designed agents may bear no relevant differences to those exemplars[10]. However, in the eyes of the would-be designer of a responsible agent,

[9] Tierney & Glick offer several further technical challenges about balancing different aspects of causal sourcehood, but those seem resolvable in principle.

[10] 'Hierarchical' and 'deep-self' accounts of responsibility which hold that the conditions of responsibility lie in the relationship between one's internal mental states fall into this category. The reliance on exemplars derives from the account's focus on there being 'no relevant

the resulting design constraint is that their designed agent must bear no relevant difference to those exemplars. This is challenging, because there seems always to be an open question about whether any functional difference in an agent is relevant to responsibility – that is, whether it reflects a shortcoming in the agent's agential structure. The only way to *ensure* responsibility is to design an agent that is functionally identical to a responsible agent. But this limits the design space to agents that perform the same functions as non-designed agents (a highly demanding design constraint).

On the other hand, if we take Fischer's hard-line reply that situation is quite different, and the realm of designed responsible agents can expand greatly. This reply is genuinely permissive. For Fischer, there are no constraints on an agents' design so long as that agent is appropriately reasons-responsive – that is, so long as it acts for reasons that represent its interests relatively consistently and effectively.

One concern here is that if all an agent requires to be responsible is to be appropriately reasons-responsive, then some designed agents can take responsibility even when they are designed to defer, have no control over their values, or have otherwise idiosyncratic awareness of reasons (Gudmunsen, 2024; Hills, 2009; Purves et al., 2015). While this remains unintuitive to me, designing an agent to be reasons-responsive is certainly an achievable aim for agent-designers and offers a wide design space.

Soft-line replies entail a stronger relationship between design and responsibility. On these accounts, design often undermines a designed agent's responsibility because design must be performed by an agent and assign a predictively effective function, both of which imply that the designer must 'effectively intend' a designed agent to act and that the designer is the 'causal source' of the designed agent's actions. If so, then under the soft-line reply no designed agent can be responsible, as Hakli & Mäkelä claim.

Filippo Santoni de Sio & Jeroen van den Hoven appear to support this line in their interpretation of Fischer's 'ownership' condition for responsibility. They suggest that designed agents are either part of a wider socio-technical system or in a state analogous to manipulation:

> Systems whose actions and states are not traceable to relevant understanding and endorsing by some human person—be they a designer, a controller, a user, etc.—no matter how intelligent and reason-responsive they may be, are not under meaningful human control. They would be like human actions carried out under psychological manipulation, subliminal persuasion, brainwashing, and indoctrination; here, the agent's behavior is clearly responsive to someone's reasons, but not to the agent's reasons. (Santoni de Sio & van den Hoven, 2018, p. 9)

Unlike Fischer, Santoni de Sio & van der Hoven imply that designed agents do *not* 'own' their own actions. They support the idea that design is manipulative through something like a 'causal-sourcehood' account – artificial agents cannot be their own 'causal sources', and so the only way to resolve responsibility gaps is to ensure that artificial agents are under 'meaningful human control', i.e., that there is some responsible human who can take ownership over the action.

However, even if soft-line replies are correct, I believe it is still possible to design responsible agents. Designing responsible agents requires designing reasons-responsiveness or functional equivalence (as in hard-line replies), and additionally designing around conditions concerning causal sourcehood (possibly including a condition causal independence from other agents).

While the addition of further conditions may appear to be more restrictive. In practice they may be less restrictive than Frankfurt/Watson's accounts. Ironically, the addition of a causal sourcehood condition may lift some of the burden on functional equivalence. Satisfying conditions for causal sourcehood permits less emphasis on functional equivalence, because an agent only loosely functionally equivalent to a responsible agent can be differentiated from non-responsible agents by being its own causal source.

However, the design constraints imposed by causal-sourcehood conditions are more imposing than for those entailed by Fischer's reasons-responsiveness account. Including causal sourcehood conditions does mitigate concerns about deferential or other reasons-responsive but intuitively non-responsible agents, but this comes at a cost. The challenge is to achieve reasons-responsiveness without a designer being the causal source of the designed agent's actions. In effect, the requirement here is that the causal influence of the designer is 'laundered' such that the designer, while able to adequately intend and plan out a broad function for the designed agent (such that the definition of design is satisfied), is not the causal source of most of the designed agent's actions (and nor is any other agent).

In practice, two ways of doing so would be to introduce 'randomness' and create a 'random agent' (Musiał, 2022, Sect. 5), or to lean on game-theory and evolutionary forces and create an 'evolutionary agent' (such as in 'artificial life'). But whatever method used to meet the design constraint of 'laundering' one's causal influence on designed agents, there seems to be a plausible design space for designing responsible agents even if the soft-line reply is correct.

difference' between designed and responsible agents' mental state structure.

Conclusion

Positions in machine ethics vary between the claim that no designed agent can be responsible (Hakli & Mäkelä) and the claim that design is irrelevant to responsibility (Behdadi and Munthe). These positions have counterparts in discussions of design cases in the literature on responsibility. Hakli & Mäkelä support a soft-line reply such that design violates a condition of responsibility; while Behdadi & Munthe support a hard-line reply such that an agent's origin cannot violate conditions of responsibility. I have also identified further positions such as Fischer's position that responsibility has an ownership condition that is not violated by design (a variety of hard-line reply) and Santoni de Sio & van den Hoven's position that responsibility has an ownership condition that *is* violated by design (a variety of soft-line reply).

I have suggested that under all the mainstream theories of responsibility, it is possible to design a responsible agent. However, doing so may carry onerous conditions (the soft-line reply) or it may seem overinclusive (Fischer's hard-line reply) or hard-to-apply (Frankfurt/Watson's hard-line reply). One clear take-away from this is that the relationship between design and responsibility is complex – while there isn't a simple condition such as 'no designed agent can be responsible', there are several potential conditions of responsibility that a designer must carefully consider when intending to design a responsible agent. Depending on your favoured reply to design cases, those constraints can vary significantly. It will be easiest to design a reasons-responsive agent under a hard-line reply to design cases, but even in this case there is likely to be disputes about whether the designed agent *truly* responds to reasons. Satisfying design constraints under soft-line replies is likely to be more difficult because it will involve 'laundering' one's causal influence while still achieving reasons-responsiveness or (a level of) functional equivalence to some responsible agent. However, as I've suggested, this appears to be *easier* than simply designing an agent to be functionally equivalent to a responsible agent alone (as the Frankfurt/Watson hard-line reply would suggest), because a causal sourcehood condition can reduce conditions for the closeness of functional equivalence required.

Designing responsible agents represents a theoretical avenue for resolving responsibility gaps – if designed agents can take responsibility for their own actions, then they can perform complex tasks without incurring moral concern from humans. But work in understanding the workings of sociotechnical systems with non-responsible artificial agents, including how we ought to assign responsibility within them, is not going to be eclipsed by this. Responsible and non-responsible designed agents are suited to carrying out different roles and functions. The use-cases for designed responsible agents is small but significant (and with a lot of potential to grow).

The prospects of using random and evolutionary forces to design responsible agents (and indeed fully fledged moral agents) to avoid the charge of manipulative design are worth exploring further, especially in relation to how well they can satisfy causal-sourcehood accounts of responsibility. I have only argued here for the possibility of designing responsible agents. I have also identified some of the potential design constraints for doing so, but satisfying those constraints while designing full moral agents is a further challenge that I have not explored here.

Acknowledgements Thanks to those who gave valuable contributions and feedback when I presented earlier versions of this paper at Bilkent University, Ethics and Technology Early-Career Group, and the IDEA centre, University of Leeds. Particular thanks to Gabriela Arriagada-Bruneau, Helen Beebee, Graham Bex-Priestley, Michael Cannon, John Danaher & Rob Lawlor for their comments on earlier drafts.

Author contributions Zacharus Gudmunsen is the sole author of this work.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ayer, A. J. (1963). Freedom and Necessity. In A. J. Ayer (Ed.), *Philosophical Essays*. Springer.

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*, *30*(2), 195–218. https://doi.org/10.1007/s11023-020-09525-8

Borgo, S., Franssen, M., Garbacz, P., Kitamura, Y., Mizoguchi, R., & Vermaas, P. E. (2014). Technical artifacts: An integrated perspective. *Applied Ontology*, *9*(3–4), 217–235. https://doi.org/10.3233/ao-140137

Bringsjord, S. (2008). Ethical Robots: The Future can heed us. *AI and Society*, *22*(4), 539–550. https://doi.org/10.1007/s00146-007-0090-9

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G. J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics*, *3*(1), 241–255. https://doi.org/10.1007/s43681-022-00167-3

Chomanski, B. (2019). What's wrong with designing people to serve? *Ethical Theory and Moral Practice*, *22*(4), 993–1015. https://doi.org/10.1007/s10677-019-10029-3

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299–309. https://doi.org/10.1007/s10676-016-9403-3

Danaher, J. (2020). Welcoming Robots into the Moral Circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, *26*(4), 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Danaher, J. (2022). Tragic choices and the Virtue of Techno-responsibility gaps. *Philosophy & Technology*, *35*(2), 26. https://doi.org/10.1007/s13347-022-00519-1

De Marco, G. (2023). Manipulation, machine induction, and bypassing. *Philosophical Studies*, *180*(2), 487–507. https://doi.org/10.1007/s11098-022-01906-2

de Santoni, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, *5*. https://www.frontiersin.org/articles/https://doi.org/10.3389/frobt.2018.00015

Deery, O., & Nahmias, E. (2017). Defeating manipulation arguments: Interventionist causation and compatibilist sourcehood. *Philosophical Studies*, *174*(5), 1255–1276. https://doi.org/10.1007/s11098-016-0754-8

Dennett, D. C. (1984). *Elbow room: The varieties of Free Will Worth Wanting*. MIT Press.

Fischer, J. M. (2011). The Zygote argument remixed. *Analysis*, *71*(2), 267–272.

Fischer, J. M. (2014). Review of Free will, agency, and meaning in life, by Derk Pereboom. *Science Religion and Culture*, *1*, 202–208.

Fischer, J. M. (2016). How do manipulation arguments work? *The Journal of Ethics*, *20*(1), 47–67. https://doi.org/10.1007/s10892-016-9225-x

Fischer, J. M. (2017). Responsibility, autonomy, and the Zygote argument. *The Journal of Ethics*, *21*(3), 223–237.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press. https://doi.org/10.1017/CBO9780511814594

Frankfurt, H. G. (1988). *The importance of what we care about: Philosophical essays*. Cambridge University Press.

Frankfurt, H. G. (2002). Reply to Fischer. In S. Buss, & L. Overton (Eds.), *Contours of Agency: Essays on themes from Harry Frankfurt*. MIT Press.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology*, *10*(2), 115–121. https://doi.org/10.1007/s10676-008-9163-9

Gudmunsen, Z. (2024). The moral decision machine: A challenge for artificial moral agency based on moral deference. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00444-3

Hakli, R., & Mäkelä, P. (2016). Robots, autonomy, and responsibility. In J. Seibt, M. Nørskov & S. S. Andersen (Eds.), *What social robots can and should do: Proceedings of robophilosophy 2016* (pp. 145–154). IOS Press.

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

Hellström, T. (2013). On the Moral responsibility of Military Robots. *Ethics and Information Technology*, *15*(2), 99–107. https://doi.org/10.1007/s10676-012-9301-2

Hills, A. (2009). Moral Testimony and Moral Epistemology. *Ethics*, *120*(1), 94–127. https://doi.org/10.1086/648610. JSTOR.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*(1), 19–29. https://doi.org/10.1007/s10676-008-9167-5

Ioan, M., & Howard, D. (2017). Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency. In T. Powers (Ed.), *Philosophy and Computing: Essays in epistemology, philosophy of mind, Logic, and Ethics*. Springer.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*(4), 195–204. https://doi.org/10.1007/s10676-006-9111-5

Kearns, S. (2012). Aborting the Zygote argument. *Philosophical Studies*, *160*(3), 379–389. https://doi.org/10.1007/s11098-011-9724-3

King, M. (2013). The Problem with Manipulation. *Ethics*, *124*(1), 65–83. https://doi.org/10.1086/671391

Köhler, S., Roughley, N., & Sauer, H. (2018). Technologically blurred accountability? Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Debiel (Eds.), *Moral Agency and the Politics of Responsibility*. Routledge.

Latham, A. J., & Tierney, H. (2022). Defusing existential and universal threats to Compatibilism: A Strawsonian Dilemma for Manipulation arguments. *Journal of Philosophy*, *119*(3), 144–161. https://doi.org/10.5840/jphil202211939

Liu, X. (2022). Manipulation and machine induction. *Mind*, *131*(522), 535–548. https://doi.org/10.1093/mind/fzaa085

Matheson, B. (2016). In defence of the four-case argument. *Philosophical Studies*, *173*(7), 1963–1982. https://doi.org/10.1007/s11098-015-0587-x

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning Automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McKenna, M. (2008). A hard-line reply to Pereboom's four-case Manipulation Argument. *Philosophy and Phenomenological Research*, *77*(1), 142–159.

McKenna, M. (2016). A modest historical theory of Moral responsibility. *The Journal of Ethics*, *20*(1/3), 83–105.

McLaughlin, P. (2000). *What functions explain: Functional explanation and self-reproducing systems*. Cambridge University Press. https://doi.org/10.1017/CBO9780511498510

Mele, A. R. (2008). *Free Will and Luck* (1st edition). Oxford University Press.

Mele, A. R. (2013). Manipulation, Moral responsibility, and bullet biting. *The Journal of Ethics*, *17*(3), 167–184. https://doi.org/10.1007/s10892-013-9147-9

Mickelson, K. (2015). The Zygote argument is invalid: Now what? *Philosophical Studies*, *172*(11), 2911–2929. https://doi.org/10.1007/s11098-015-0449-6

Millikan, R. G. (1999). Wings, spoons, pills, and Quills: A pluralist theory of function. *The Journal of Philosophy*, *96*(4), 191–206. https://doi.org/10.2307/2564702

Musiał, M. (2017). Designing (artificial) people to serve – the other side of the coin. *Journal of Experimental & Theoretical Artificial Intelligence*, *29*(5), 1087–1097. https://doi.org/10.1080/0952813X.2017.1309691

Musiał, M. (2022). Can we design artificial persons without being manipulative? *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01575-z

Oimann, A. K. (2023). The responsibility gap and LAWS: A critical mapping of the debate. *Philosophy & Technology*, *36*(1), 3. https://doi.org/10.1007/s13347-022-00602-7

Pereboom, D. (2001). *Living without Free Will*. Cambridge University Press. https://doi.org/10.1017/CBO9780511498824

Petersen, S. (2011). Designing people to serve. In P. Lin, G. Bekey, & K. Abney (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 283–298). MIT Press, Cambridge.

Powers, T. M. (2013). On the Moral Agency of Computers. *Topoi*, *32*(2), 227–236. https://doi.org/10.1007/s11245-012-9149-4

Preston, B. (2009). Philosophical theories of artifact function. In A. Meijers (Ed.), *Philosophy of Technology and Engineering Sciences* (pp. 213–233). North-Holland. https://doi.org/10.1016/B978-0-444-51667-1.50013-6

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous Machines, Moral Judgment, and acting for the right reasons. *Ethical Theory and Moral Practice*, *18*(4), 851–872. https://doi.org/10.1007/s10677-015-9563-y

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy and Technology*, *26*(2), 203–219. https://doi.org/10.1007/s13347-012-0089-0

Sullins, J. P. (2006). When is a robot a moral agent. *International Review of Information Ethics*, *6*(12), 23–30.

Tierney, H., & Glick, D. (2020). Desperately seeking sourcehood. *Philosophical Studies*, *177*(4), 953–970. https://doi.org/10.1007/s11098-018-1215-3

Torrance, S. (2008). Ethics and consciousness in Artificial agents. *AI and Society*, *22*(4), 495–521. https://doi.org/10.1007/s00146-007-0091-8

Usher, M. (2018). Agency, teleological control and robust causation. *Philosophy and Phenomenological Research*, *100*(2), 302–324. https://doi.org/10.1111/phpr.12537

Vallor, S., & Vierkant, T. (2024). Find the gap: AI, responsible Agency and Vulnerability. *Minds and Machines*, *34*(3), 20. https://doi.org/10.1007/s11023-024-09674-0

Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & SOCIETY*, *36*(2), 487–497. https://doi.org/10.1007/s00146-021-01189-x

Veluwenkamp, H. (2022). Reasons for Meaningful Human Control. *Ethics and Information Technology*, *24*(4), 51. https://doi.org/10.1007/s10676-022-09673-8

Vermaas, P. E., & Houkes, W. (2006). Technical functions: A drawbridge between the intentional and structural natures of technical artefacts. *Studies in History and Philosophy of Science Part A*, *37*(1), 5–18. https://doi.org/10.1016/j.shpsa.2005.12.002

Vermaas, P. E., Carrara, M., Borgo, S., & Garbacz, P. (2013). The design stance and its artefacts. *Synthese*, *190*(6), 1131–1152.

Walker, M. (2006). A moral paradox in the creation of artificial intelligence: Mary Poppins 3000s of the world unite! In T. Metzler (Ed.), *Human implications of human-robot interaction: Papers from the AAAI workshop* (pp 23–28). AAAI Press, Menlo Park.

Waller, R. R. (2014). The threat of effective intentions to Moral responsibility in the Zygote argument. *Philosophia*, *42*(1), 209–222. https://doi.org/10.1007/s11406-013-9476-8

Watson, G. (1999). Soft libertarianism and hard compatibilism. *The Journal of Ethics*, *3*(4), 351–365. http://www.jstor.org/stable/25115625

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, *91*(2), 303–347. https://doi.org/10.1111/phpr.12095

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.