Minds and Machines (2023) 33:525–539
https://doi.org/10.1007/s11023-023-09643-z

Avatars as Proxies

Paula Sweeney[1]

Received: 23 March 2023 / Accepted: 9 July 2023 / Published online: 26 July 2023

Abstract

Avatars will represent us online, in virtual worlds, and in technologically supported hybrid environments. We and our avatars will stand not in an identity relation but in a proxy relation, an arrangement that is significant not least because our proxies' actions can be counted as our own. However, this proxy relation between humans and avatars is not well understood and its consequences under-explored. In this paper I explore the relation and its potential ethical consequences.

Keywords Avatars · Proxies · Avatar representation · Responsibility gaps

The purpose of this paper is to explore the avatar-as-proxy relation and its potential consequences. The paper provides a framework for thinking about the human-avatar proxy relation, drawing on our understanding of the nature of traditional proxies and expanding on distinctions regarding the general features of online proxies. Of particular interest are the conditions under which a person can reasonably be held responsible for an avatar proxy's actions and the potential consequences of such an arrangement.

There are a number of philosophical debates which, although nearby, are distinct from the questions addressed in this paper. There has been much recent discussion around the possibility of uploads. An 'upload' is an entity that has transferred its thoughts, memories and characteristics from an organic brain onto a computer. It might then have a virtual or simulated body or could exist in an android body. Uploading leads to questions of personal identity. Is uploaded Adam the same person as organic Adam? Or, even if they are not identical, is the upload a continuation

✉ Paula Sweeney
p.sweeney@abdn.ac.uk

[1] University of Aberdeen, Aberdeen, UK

of Adam, preserving some of his aspects?[1] In this paper I am focussing on similar but less technologically distant extensions of existing technologies through which humans will employ avatars, not as continuations or replications of themselves, but as proxies. Virtual reality experiences, telepresence systems and 'continuing bond' representations of deceased loved ones are all evidence of our desire to explore advancing technology that will allow a human to be represented in a different environment. In using a proxy the represented person will not themselves *be* in the different environment because proxies are by definition entities that are distinct from the things that they represent. However, as the proxy relation is one in which the represented person can be held responsible for the actions of their proxy, there are distinct ethical considerations arising from such arrangements.

This is reminiscent of another nearby philosophical debate regarding responsibility gaps for AI-determined actions.[2] That debate moves around the question of who can be held responsible for unwelcome consequences when there is no obvious candidate to be fairly identified as the responsible ethical agent. In this paper I am exploring a different AI responsibility problem, one that arises from an epistemic gap between the represented agent and their avatar proxy. As expanded below, when a non-degenerate proxy relation is in place, there is a clear line of responsibility; the represented person is responsible for their proxy's action. In that sense, the familiar responsibility gap problem does not arise. However, I argue that avatar proxy relations come with a distinctive epistemic gap which potentially exposes represented agents to a high risk of being personally responsible for actions that they would not themselves perform. This is because an avatar proxy will sometimes be required to make a decision or perform an action without knowing what the represented agent would do in that situation. This, combined with the fact that an agent is responsible for the actions of their proxy, leads to an epistemic gap responsibility problem. With no reliable mechanism for filling the epistemic gap, the use of avatar proxies will create a distinctive problematic responsibility relation.

My ultimate aim is to demonstrate the risks that arise from establishing human-proxy avatar relationships. In Sect. 1 I consider the conditions under which a proxy relation might be desirable, examine the features of such a relation and the responsibilities that arise. In Sect. 2 I take a closer look at avatars with a view to determining what kinds of avatars might be considered human proxies and under what conditions. In Sect. 3 I identify a distinct responsibility risk that arises from the use of avatar proxies. I focus on what I call the proxy epistemic gap and argue that the responsibility relation under which agents are responsible for the actions performed by their proxies is particularly problematic when the proxy is an avatar.

[1] For more on uploads see, for example, Corabi and Schneider (2012), Chalmers (2014, 2022) and Walker (2014). I do not focus on uploads in this paper because here I am explicitly interested in problems that arise when the relation is not identity nor even 'copy' but is a proxy relation. One notable difference between a proxy and an upload is that in the proxy case the represented agent and their avatar must be distinct from each other whereas in the case of uploads efforts are made to show how the upload is identical to the original. I discuss uploads and identity in Sweeney (forthcoming).

[2] See, for example, Nyholm (2017) and Coeckelbergh (2019).

1 Proxies

Proxy arrangements are not new. Traditionally a proxy is understood as a person with the authority to act on behalf of another. For example, you might be appointed to make financial decisions on behalf of an elderly parent, making you their proxy in financial matters. But the proxy relation is not limited to humans. Online entities act as proxies for businesses to enable them to interact with customers in a digital environment.

In this paper I am interested in exploring the proxy relation as it might develop between humans and avatars: the nature of the relation, the duties of each party, and the risk to the represented agent if things don't go according to plan.

1.1 The Proxy Relation

What is the proxy relation? In simple terms, a proxy is an entity that stands for another entity. However, not everything that 'stands for' is a proxy in the full sense of the word. Floridi (2015) notes that acting by proxy is possible because something both represents *and* replaces something else. Floridi calls this the 'vicarious' relation.

Proxy: P is a proxy for R if and only if P both 'stands for' R and can 'stand in for' R.

According to Floridi, two 'degenerate' proxy relations follow: *signs* and *surrogates*.

> If P has a vicarious relation with (i.e. it is a proxy for) R with a zero degree of 'standing in for' R, P is a degenerate proxy that only stands for, but cannot behave on behalf, or act instead, of R. Such degenerate proxies are *signs*. Similarly, if P has a vicarious relation with R but now with a zero degree of 'standing for' R, then P is another kind of degenerate proxy, one that only behaves on behalf, or acts instead, of R without referring to it. Such degenerate proxies are *surrogates*. (2015: 488)

Floridi notes a possible further division of the first kind of degenerate proxies, sign-proxies, into icons, indexes and symbols but for our purposes it will be enough to note that a sign-proxy is something that signifies another entity and cannot stand in for it. A good example of this kind of proxy relationship is that between a word and the object that it signifies. A word can stand for the object that it signifies but it cannot stand in for it—you cannot take a drink from the word 'cup'.

The second kind of degenerate proxy relation is a proxy that acts on behalf of something but cannot stand for it. For example, oat milk is often used as a replacement for cow's milk but it does not 'stand for' or represent cow's milk. This is a *surrogate-proxy*.

Floridi states:

> […] proxies are pragmatically more than signs because they are signifiers that also stand in for the signified and so you can interact with them instead of interacting with the signified. And they are epistemologically more than surrogates,

> because they are signifiers you can interact with that refer to the signified they replace, so you can still perceive the difference. (2015: 489)

Floridi's paper is insightful but brief and the distinctions are subtle. Floridi does not himself discuss the potential proxy relation as it might arise between a human and an avatar. In the rest of this section, I develop Floridi's framework and use it to make sense of potential avatar-human proxy relationships.

In identifying whether a proxy relation is degenerate or non-degenerate it helps to consider who is deemed responsible when the proxy performs an action. For, as Floridi notes, it is only when the proxy relationship is non-degenerate that the actions of the proxy are considered to be the actions of the represented agent. The two degenerate proxies fall short of this in different ways. The sign-proxy is not taken to stand in for the agent at all, so the actions of a sign cannot be understood to be the actions of the represented agent. The surrogate-proxy falls short in a different way. The surrogate stands in for you, perhaps in the sense of performing some action that you require, but the link that connects you and the surrogate fails. Floridi does not make this observation but it seems plausible that the link can fail to obtain in (at least) two ways. The link might never have been there in the first place. For example, if you ask someone to keep your place in line, they are standing in for you but do not stand for you. You could not reasonably be held responsible for their actions. Second, it seems possible that a non-degenerate proxy relation might deteriorate into a degenerate proxy relation if the community forgets or rejects the signifying aspect of the relationship. Imagine, for example, that Bill can no longer make community council meetings and asks John to attend as his proxy. Initially the council members see John as standing for and standing in for Bill. Over time, despite the fact that John faithfully acts in ways that represent Bill, the community council interact with John as John and think of him no longer as Bill's representative but as an agent in his own right. In such a case it seems that the proxy relation could begin to fail. The risk arises particularly in situations where, over time, we stop being able to perceive the difference between the proxy and the thing it stands for. (Floridi, 2015) When that referencing relation fails, the ability to reasonably hold the agent responsible for the proxy's action can fail with it. That this is possible highlights another significant feature of the proxy relationship, that it is not a strict contractual relationship but a pre-legal arrangement, established by mutual consent and maintained by community recognition. (MacPherson, 2010)

1.2 Responsibilities of the Proxy Relation

What are the responsibilities of the proxy and the represented agent? [3] Millar (2014) notes that the ideal proxy-decision is the one that the agent would choose if they were in a position to make the judgement themselves. In fact, there is a sense in which the ideal proxy decision-maker is not making a decision at all—they are simply channelling or acting as a conduit for the decision of the agent. For example, if you are my designated healthcare proxy and I have made it clear to you that I would like to

[3] From here onwards I use 'proxy' to mean non-degenerate proxy.

donate my organs if anything happens to me, to fulfil the role of proxy it is your duty to make my wishes known, regardless of your own views on organ donation.

But the reality is that often proxy decision-makers do not have all of the information that they would need in order to reach the decision that the agent would make. For example, you may be my medical proxy but at the time of an accident we have never discussed my organ donor preferences, so you are expected to somehow fill the gap in your knowledge. However, even in such cases, as proxy you must remove your own preferences from the decision-making. You are expected to think about what my organ donation preferences might be, given other things that you know about me, or to look for other supporting evidence.

Another feature of a proxy arrangement is that it usually has limits placed on it. Consider the situation in which Bill has chosen John as a proxy to vote on his behalf at a local council meeting. While attending the meeting John starts to pick a fight with another attendee. When he casts the vote John's actions counts as Bill's actions. Moreover, Bill is responsible for the consequences of those actions. However, this does not extend to John's general behaviour at the meeting: Bill could not reasonably be held responsible for John's violence, nor for the consequences of that violence. Despite John being Bill's proxy at the meeting, the proxy arrangement starts and ends with the casting of the vote. These limits of the proxy relationship between John and Bill may have been explicitly drawn but it is more likely that they were implicit, made clear by the history of proxy voting arrangements and the context of the agreement.

In Sect. 3 I consider how these features—the lack of clearly defined limits to the proxy arrangement and the epistemic gap that proxies will often face—might play out in the case of avatar proxies. But first, what kinds of avatars are likely to be relevant to a proxy relationship?

2 Avatars

The possibility of an advanced avatar that provides an acceptable representation of a person in a full proxy relation is perhaps still some way off. However, proxy relations are established and maintained by social convention and, depending on what standards for proxy representation we are willing to accept, the possibility is not so distant that we can be complacent.

At-a-distance 'in-person' interactions already exist, albeit with a fairly direct link between the represented person and the representing system. Consider telepresence systems. These are intended to allow a person to feel as if they were present at a different location and perhaps to give the appearance to those they interact with via the system that they are present there. The system allows the user to move a camera and speak with a microphone through a speaker, often with a display of the represented person's face. Alongside their potential use in industries such as tourism and education, they can be used to improve the lives of people who have special needs, such as those who are bed-bound, and they can allow for physically distant people to 'visit' those they care about. This technology permits a robot to physically represent a person; it is a sign proxy. Using Floridi's terminology, it stands for but is not required to stand in for or act on behalf of in any interesting sense because it is directly controlled by the person it stands for. However, it is not a big leap to imagine that we might desire a system in which the entity that stands for the person can stand in for them too.

Technology will advance to such a stage that we are likely to allow for representative systems that do not require our direct control while acting on our behalf. After all, proxies are often required precisely when we are looking for someone to stand in for us because we are unavailable.

What kinds of avatars might be considered to be human proxies and under what conditions? What an avatar is, the capabilities and the potential benefits they bring, is changing with both advances in technology and with our understanding of the conditions under which something can be thought of as representing someone. The taxonomy below is in no way exhaustive and, as I explain below, avatars are difficult to classify neatly, but it does give us some cases to hang our imaginings of future possibilities on.

Memoji Avatar The most basic form of avatar we might be familiar with is a memoji; an emoji that is designed to look something like the person it represents. It has only basic interactive abilities.

Gaming Avatar This is perhaps the most common modern use of the term ‘avatar’. In early computer games players were encouraged to identify with the player characters, seeing the characters as representations of themselves in the gaming environment. As gaming technology has advanced, players have been able to design the avatars that represent them. This creates a closer emotional connection between the player and their avatar and has been shown to further increase player engagement.[4]

Advanced Memoji Avatar In 2020 Kanye West gifted Kim Kardashian a hologram of her dead father to wish her happy birthday. I view this hologram as another form of memoji, albeit an advanced one. It uses fine-grained information about a particular person to create an advanced representation of them. The Kardashian example, created using images and voice recordings of her father, is just one example; there are others of varying degrees of sophistication. Chatbots such as ‘DadBot’, built from interviews with the creator James Vlaho’s deceased father, and ‘Roman’ built on a neural network that was fed with Russian text and thousands of telegrams exchanged between Eugenia Kuyda’s friends and their deceased friend Roman are further examples.[5] I see these as extensions of the memoji form as they take a static snapshot of information and create a detailed representation of a person with it.

We can imagine a dynamic version of the advanced memoji, call it the *Autonomous Advanced Avatar*, that could dynamically represent a living person who is not in the environment that the system is operating in. There are various ways of padding out the dynamic nature of the system. For example, we can imagine an avatar that operates something like ‘Dadbot’ but, rather than having a closed set of data, it could

[4] See Banks and Bowman (2013).

[5] See Krueger and Osler (2022).

utilise a growing set of data derived from the ongoing interactions of the individual it represents—a growing set of dialogue, evidence of preferences and decisions made etc.—and use that to build a predictive model of behaviour. The model could be the basis for an avatar version of that person.[6] However we fill out the details, we can certainly imagine models in which an avatar can represent you without you having to be cognitively present as it does so.

Returning now to Floridi's distinction, although all of the entities described above satisfy some form of the proxy relation, the autonomous advanced avatar provides us with our clearest case of a non-degenerate proxy. Memojis, and even Advanced Memojis, are arguably sign-proxies. Memojis perform a similar function as a name or a picture. They *stand for* a person but do not *stand in for* a person. Even the Advanced Memoji does not straightforwardly stand in for a person (although, see the notes on the context-sensitive nature of 'standing in for', below.) It is a hologram or chat bot that has been constructed using a fixed data set of information from past interactions, a very advanced picture of a person as they were.

Gaming Avatars, as they currently are, clearly do stand in for or represent the player in the gaming environment as surrogate proxies, but do they signify or refer to the player? They certainly have the potential to do so within the fictional environment of the game. Players will very often refer to their player-character in the first person (e.g. "I can't believe I just killed all of those aliens!") and when playing a multiplayer game others might legitimately refer to the player-character as the player, informing other players that, for example, "Josh is 'Wildcat'". On the other hand, the fictional element complicates things with the result that it is not clear that Wildcat is a full proxy for Josh. Certainly Josh is not considered morally responsible for the actions of Wildcat in the way that he would be considered responsible for similar real world actions.[7] However, as players become more involved in the individual design of their avatars and begin to consider that they are playing online games *as themselves*, and as the gaming environment starts to realistically represent the real world, the distance between gaming avatars and autonomous advanced avatars is likely to reduce substantially such that gamers may well become fully responsible for the actions of their gaming characters.

The autonomous advanced avatar is a clear example of an avatar that can both stand for and stand in for an agent: it is a non-degenerate proxy. It combines the sign-proxy features of the advanced memoji with the surrogate proxy features of a gaming avatar.

It is useful for me to identify a clear instance of a non-degenerate avatar proxy but in reality, what counts as a non-degenerate or full avatar proxy can change for two reasons. The first reason is that the meeting of the two conditions that form the full proxy relation, standing for and standing in for, is partly socially determined and context sensitive. Consider the example of 'Dadbot' given above. There I said

[6] Of course the avatar is not entirely autonomous because what it chooses to do is heavily influenced by your preferences and goals. It is autonomous in the sense that it is able to make decisions without direct instruction.

[7] See Luck (2009) and Davnall (2021) for just some of the large body of discussion on player responsibility.

that Dadbot could stand for Vlaho's father but could not stand in for him; Dadbot is a sign-proxy but does not fulfil the surrogate condition so is not a full proxy. But we can imagine a situation in which it is plausible that a 'continuing bonds' robot, a robot that uses a closed set of data from a deceased person's life, is counted by some people as standing in for the deceased person, thereby arguably fulfilling the full proxy conditions. In fact, this very scenario is represented for us in Charlie Brooker's *Black Mirror* episode, 'Be right back'. (Brooker, 2016) The episode depicts a scenario in which a recently bereaved woman, Martha, uses her deceased partner's (Ash) online communications to create a messaging service based on him. She gradually creates more advanced versions of Ash, first interacting with a video version of him until finally creating an identical android. As the episode progresses, it is clear that Martha oscillates between considering the entity she is interacting with as sometimes fulfilling the standing in for relation, and other times as falling short of fulfilling it. That is, it seems that the exact same entity can at sometimes and by some people be counted as fulfilling the full proxy relation and at other times not.[8] The second shiftiness is tied more closely to how technological developments will allow us to fulfil the proxy relation. It is this shiftiness that explains potential variability with the example of gaming avatars. We do not currently see gaming avatars as full proxies for the humans that they represent, as evidenced by the fact that we do not hold humans responsible for their avatars' actions. But at some point, either when the avatar draws more closely on the characteristics of the represented human or when the gaming world more closely represents the actual world, we will see gaming avatars as full proxies.

This shiftiness of full avatar proxy status is not simply an academic complication, it is a serious consideration with significant real-world repercussions. Given the responsibility condition that comes with a full proxy relation the variability and unclarity around what counts as a fulfilment of the full proxy condition is a concern. In particular, the gradual technological developments that could push an avatar from having degenerate proxy status to having non-degenerate proxy status without warning is a consideration worthy of our immediate attention.

3 Avatar Proxies

What would motivate the use of something like an autonomous advanced memoji avatar as a proxy? The most obvious motivation is that the avatar allows the agent to interact in a different environment. Mark Zuckerberg's launch of the Metaverse drew much mockery because the environment did not look like somewhere that anyone would want to spend time, but that will change and undoubtedly there will soon be digital environments that we will want to interact in. Another motivation is that the use of avatars can open up experiences that might be otherwise closed to us. For

[8] This is in some ways reminiscent of a discussion in the literature around personal identity. Nozick (2014) was moved to introduce the Closest Continuer theory of personal identity to accommodate the fact that personal identity conditions can vary; under some conditions we intuitively favour physical continuity over psychological continuity and visa versa. That is, there could be a social element to both personal identity and proxy conditions.

example, the internet can provide an accessible alternative for people with disabilities to activities that usually require physical mobility. (Garaj, Dudley and Kristensson, 2022) In virtual environments, users with disabilities would represent themselves using avatars and access social interactions, education and even work that might otherwise be inaccessible to them.[9]

As noted above, without human-uploading technology, our interactions in virtual worlds will necessarily be 'once removed', via avatar proxies. When the motivation for using a proxy is to have a new experience, such as in Zuckerberg's Metaverse, we are likely to want to have fairly direct real-time control over our avatars. Without that presence, we wouldn't be able to feel like we are experiencing a new environment at all. Because of this it might seem that there would be little to be gained from sending your avatar off to play in the Metaverse while you go to work. But if we think beyond the gaming experience, there may well be reasons to do just that. If the Metaverse, or some future variation of it, gave opportunities for other social goods such as the creation of digital wealth or social status, then we may well send our avatars into the digital environment to 'work' on our behalf.[10]

And it is certainly conceivable that there will be a future in which we will send an avatar proxy to a virtual meeting to act on our behalf. The obvious motivation here is that humans are limited to being in one place at a time. Rather than delegate attendance to a deputy, an overstretched executive might well place trust in their avatar proxy, allowing the executive to perform actions and be represented in multiple meetings at once. Another motivation we can imagine for remaining related to, but distinct from, the experiences of our avatar proxies is if we might be responsible for performing some action that we are likely to find difficult. For example, a doctor might find it distressing to deliver bad medical news to a patient and, all things being equal, would prefer if that news could be delivered by an avatar proxy, meeting their responsibility while lightening the emotional burden of their role.[11] Or a person might be excellent at all other aspects of their role but be really bad at giving public presentations and prefer to delegate that part of the job to their avatar proxy.

These are just some of the motivations for humans to have avatar proxies. But what are the potential consequences of such arrangements?

[9] The UKRI-funded project "Towards an Equitable Social VR' investigates the potential of virtual reality platforms to bring life-changing benefits to people with disabilities and older people. https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/W025698/1.

[10] Hanson (2016) describes a future society in which robots with brain emulations will displace humans in most jobs. Hanson focusses on robots with mind uploads (not proxies) and explicitly does not focus on the role that humans will play in such a world. I am not here imagining Hanson's dystopian future, but his imagining of how robots with human-like brains can be disrupters does motivate the possibility that avatar proxy workers will have a role to play in future economies.

[11] See Togni et al. (2021) for a discussion of artificial emotional intelligence in the healthcare setting, encouraging us to move away from a 'humanistic' paradigm towards a future in which robots have social emotional capacity and play a broader role in healthcare.

3.1 Indirect Harms Arising from Avatar Proxies

There are some risks that arise indirectly through having an avatar proxy. Any proxy arrangement could be *misused* to cause some kind of harm. I could manipulate your proxy to vote a different way, or your proxy could be used to purposefully misrepresent you. I noted above that the actions of the proxy will only count as those of the represented agent in defined contexts, so we might think that the reputational risk starts and ends with actions that fall within those contexts. However, it seems plausible that the very act of putting in place a proxy relationship can have a reputational consequence for the agent. In the body of literature on the question of player responsibility with regards to the actions of their gaming avatars it is sometimes proposed that any considerations of moral responsibility for the avatar's actions apply only to the actions that the player controls. However, even cut scenes—scenes in which the player character can be seen to perform acts that are *not* controlled by the player—can indirectly reflect on the moral character of the player. In the game *Custer's Revenge* there is a cut scene in which the player character Custer rapes a Native American woman. Patridge (2011) convincingly argues that playing a game with this cut scene reflects badly on the player. Patridge proposes that this is because gaining enjoyment from such a game reflects on the player's moral character. However, this explanation doesn't explain an asymmetry between playing such a game and choosing to watch a movie with a rape scene in it, which doesn't seem to reflect on the viewer's moral character in the same way. In both situations a person is voluntarily and passively watching a depiction of rape. It seems to me that what is different in the gaming example is the fact that the player has chosen to establish a (degenerate) proxy relationship with a character who would act in such a way and that the choice to establish this particular proxy relationship reflects poorly on the moral character of the player; the player is tainted by (proxy) association. This is some evidence for the fact that the very act of choosing a particular character or person to represent you can reflect on your moral character. This indicates that, contra the example of Bill and John in Sect. 1.2, a person might well be judged or tainted by the actions of their proxy *beyond* those actions that strictly fall under the proxy arrangement, purely in virtue of their choice of proxy.[12]

3.2 The Proxy Epistemic Gap and the Personal Responsibility Risk

The most significant risk of direct harm is that proxies open the represented person up to being personally responsible for actions that are not within their control. In Sect. 1 I noted two features of proxy representations. The first is that when a proxy, A, both stands for and stands in for a person, B, B is responsible for A's actions. I also described an epistemic gap that will arise when a proxy has to decide how to

[12] There is another kind of indirect harm that I do not consider here, the harm that might come to a human via the exploitation or mistreatment of their avatar. This is closely related to my discussion of indirect harms arising from robot mistreatment in Sweeney (2022). There I do not consider the distinctive case of the mistreatment of avatar proxies. A full consideration of the question of indirect harm arising from avatar abuse would also draw on the significant existing literature on avatar attachment in gaming. See, for example, Powers (2003) and Wolfendale (2007).

act without explicit instructions. I gave an example in which a proxy does not have explicit information on how to proceed but is responsible for casting a vote on the represented person's behalf. The obvious risk in such a case is that the proxy can make a decision or undertake an action that the represented agent is responsible for but that the latter had no direct part in making. Furthermore, it is left unclear how the proxy might appropriately fill the epistemic gap.

The risk that comes from the proxy epistemic gap can arise even in quite tightly constrained cases. For example, you may have given your proxy strict instructions on who to vote for at the local council election apparently eliminating any risk of misrepresentation, but you may not have specified what the proxy should do if your preferred candidate pulls out of the race. Should your proxy abstain, or should they choose another candidate for you? Or, to take another example, if your preferred candidate is exposed as a fraud right before the vote is to be cast, is your proxy right to still vote for them even if they don't know how you would respond to this updated information?

What are the duties of proxies in cases where they are faced with an epistemic gap? In particular, where is it appropriate for them to look for supporting or countervailing evidence and how should they balance these kinds of evidence? A human proxy might first try to fall back on adjacent preferences. For example, if a medical proxy knows that the person they are representing was in favour of donating their organs on their death, and they are asked whether the represented person's body could be donated to medical science, they might use the willingness for organ donation as evidence of amenability to body donation. On the other hand, if they also know that the person had wanted a traditional funeral this might be evidence to be counted against donating their body to science. The proxy has a responsibility to balance these considerations. Beyond that the proxy might seek to gather more information, perhaps by asking the represented person's friends or relatives for their opinion of what the preferences might be. But how is the testimony of different individuals to be weighted? Is testimonial evidence of more recent interactions to be given more significance than that of historic interactions?

The important point here is that, on reflection, the limits, scope and even the basic mechanics of the proxy relation turn out to be incredibly ill-defined and vague. And this is not just a feature of a lazy set up; the epistemic gap arises because the proxy relation is trying to replicate as closely as possible the mechanics of personal action and decision-making, something that is sensitive to numerous small changes in context.

The epistemic gap is a particularly significant concern in the case of avatar proxies. With human proxies the epistemic gap is left to be filled in an ad hoc, 'common sense' way. How could the epistemic gap be filled with an avatar proxy? The obvious approach would be for the avatar proxy to be enabled with some form of intelligent decision support system; the kinds of systems that are currently used to assist decision-making in areas such as finance, healthcare, marketing, commerce, and cybersecurity. These are systems that are designed to mimic human cognitive capabilities in some way. There are familiar concerns about the reliability of such systems. The high-profile cases are the ones that have made it to court such as the use of COMPAS to inform court sentencing and the use of algorithmic tools to evaluate teacher performance in some US states. (Rubel et al., 2021) But there are likely to be many more hidden examples where, for example, an automated decision system has resulted in unfair treatment to an individual through the rejection of a loan application or a decision to remove their candidacy from a pool of potential job applicants. As Rubel et el. (2021) put it, 'There is widespread recognition that there are ethical issues surrounding complex algorithmic systems and that there is a great deal of work to be done to better understand them.' (Rubel et al., 10).

Note how the situation of the avatar proxy decision maker differs from those familiar cases of algorithmic harms. In those cases, it is difficult to know where to lay the blame when things go wrong: Is the algorithm designer to blame? The people who selected the data set? The organisation who utilised the system to make decision? In the avatar proxy case, the proxy arrangement identifies the responsible agent for us. Under a proxy arrangement the represented individual will directly and personally bear the consequences of mishaps and the responsibility for any resulting harm to others. That is, we have already identified the significant negative consequences resulting from actions informed by algorithmic decision systems; the avatar proxy arrangement brings those decisions into the realm of personal action, with all of the personal responsibility that follows from it.

Even leaving catastrophic harms aside, there are less significant wrong decisions that could impact on one's reputation or, minimally, on one's sense of self. We have daily evidence of how it feels when algorithms get us wrong; the recommendations of music and books that widely miss the mark and the suggestions of things you might like to buy. Being 'misunderstood' or miscategorised by algorithms can feel like a personal insult. Imagine how it will feel when these mistakes are performed publicly by an avatar proxy that is representing you. Even small and apparently innocuous poor choices could impact on the reputation that you have with others, and perhaps ultimately on your own sense of self.

There is a further concern arising from the availability of avatar proxies. Consider again the example of using an avatar proxy in the workplace. The motivation in the example as it was described came from the employee themselves. But there may also be a sinister motivation for *employers* to encourage avatar-employee proxies. Danaher (2016) has written on the retribution gaps that can arise when an AI system causes some undesirable outcome. In the absence of a human to be held accountable for the outcome, impacted people can sometimes turn to the leaders of organisations or the owners of the relevant technology company to seek retribution. A leader who was eager to redirect responsibility for the actions of their organisations might be motivated to establish proxy relations between employees and avatar representations of them, to provide a plausible associated responsible human agent on which to hang the blame if things go wrong. That is, proxy AI systems could provide a direct route to a 'responsible' human agent to act as scapegoat, despite the epistemic gap that will exist between the agent and their proxy.

We might think that an avatar proxy could be given explicit instructions on how to act and that this would remove the risk that arises from the epistemic gap. But real-world contexts are incredibly fine-grained making it unlikely or even impossible that all eventualities can be considered in advance. As such, in order to be useful, we will need to have avatar proxies that can respond to unforeseen circumstances. Perhaps the risk could be averted if we made it that the avatar would simply not make a decision when faced with an epistemic gap? This won't work for two reasons. First, as noted above, the fine-grained nature of decision contexts means that epistemic gaps between the avatar and the represented person are likely to be ubiquitous so this restriction would severely reduce the usefulness of the proxy arrangement. Second, a decision not to act due to epistemic limitations could have its own significant consequences; that is, we can be liable for the negative consequences of acts of omission as well as those of performance. (Clarke, 2022)

As I have introduced the notion of a person's responsibility for the actions of their avatar proxy it follows from the conditions of the proxy arrangement itself; when a full (non-degenerate) proxy relation obtains, the represented person is responsible for their proxy's actions. As such, whenever we judge that a full proxy relation exists between an agent and their avatar, the responsibility condition kicks in. One might question whether this would actually be the case. Would we actually hold individuals responsible for actions that their avatar proxies perform in either real or virtual environments? I see no reason why we would not. The actions of such proxies can have real-life consequences and, as Danaher (2016) noted, in such situations we tend to look for someone to be held accountable for them; the represented agent who voluntarily entered into the proxy arrangement is the obvious 'someone'. Existing traditional proxy arrangements set precedent for such lines of responsibility. And even if we try and force a gap between the avatar proxy and the represented person, perhaps arguing that the avatar proxy representing the person is more like an employee relation than a standing-in-for relation, under the legal theory of 'respondeat superior', an employer is responsible for the acts or omissions of its employees. (Thornton, 2010) Either way, when looking for someone to hold responsible for an avatar proxy's actions, it seems inevitable that the represented person will be liable.

Clearly there are significant personal risks that arise directly from the avatar proxy relation. In order to limit these risks an ad hoc or common-sense approach to proxy arrangements will not do. We need to understand the precise limits of the proxy relation, the responsibilities of the proxy and the responsibilities of the represented agent. But increasingly sophisticated avatar proxies are being introduced without any such clarity.

The significant question we should ask ourselves is, is it possible for the loosely defined human-to-human proxy relationship to be transferred to avatar proxies in a way that doesn't expose us to significant personal risk? If not, we would be well-advised to be wary of the creeping use of avatar proxy representation.

4 Conclusion

We are moving towards a world in which humans are increasingly either acting through avatars or being represented by avatars. In many contexts the avatar will stand in a proxy relation to the represented human. Human-to-human proxy relations are already familiar to us, put in place when a person is either not available to be in a particular place at the required time or when a person is not capable of representing themselves. When the proxy relationship is in place, the actions of the proxy count as the actions of the represented agent and the represented agent can be held responsible for the consequences arising from those actions. However, proxy relations come up against epistemic gaps in which a proxy is required to act without knowing determinately what action the represented person would perform in the situation. These epistemic gaps are filled in a 'common sense', ad hoc way in a human-to-human proxy, sometimes requiring a nuanced balancing of evidence that pulls in different directions and an ability on the part of the proxy to weigh evidence from different sources. This approach, requiring sensitivity to fine-grained contexts, may be difficult to replicate with an artificial system. The potential risk is that agents are held responsible for undesirable avatar proxy actions. In sum, although avatar proxies will present opportunities for us to interact in contexts that might otherwise be difficult for us, we should be cautious when entering into human-avatar proxy relations.

Acknowledgements I am grateful to the referees for this journal who provided very helpful suggestions that greatly improved the paper.

References

Banks, J., & Bowman, N. D. (2013). Close intimate playthings? Understanding player-avatar relationships as a function of attachment, agency, and intimacy. AoIR Selected Papers of Internet Research, 3. Retrieved from https://journals.uic.edu/ojs/index.php/spir/article/view/8498.

Brooker, C. (2016). 'Be Right Back', *Black Mirror* Netflix: United States.

Chalmers, D. (2014). Mind uploading: A philosophical analysis. *Intelligence unbound: The future of uploaded and machine minds*. Blackwell.

Chalmers, D. (2022). *Reality+: Virtual Worlds and the problems of philosophy*. London: Allen Lane.

Clarke, R. (2022). Responsibility for Acts and Omissions. In D. Nelkin, & D. Pereboom (Eds.), *The Oxford Handbook of Moral responsibility* (pp. 91–136). Oxford: Oxford University Press.

Corabi, J., & Schneider, S. (2012). 'Metaphysics of Uploading', *Journal of Consciousness Studies* 19 (7).

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

Davnall, R. (2021). What does the gamer do? *Ethics and Information Technology*, *23*, 225–237.

De Togni, G., Erikainen, S., Chan, S., & Cunningham-Burley, S. (2021). 'What makes AI 'intelligent' and 'caring'? Exploring affect and relationality across three sites of intelligence and care', *Social Science and Medicine*. 277.

Floridi, L. (2015). A Proxy Culture. *Philosophy and Technology*, *28*, 487–490.

Garaj, V., Dudley, J., & Kristensson, P. O. (2022). 'Five ways the metaverse could be revolutionary for people with disabilities'. [online] *The Conversation* Available at: https://theconversation.com/five-ways-the-metaverse-could-be-revolutionary-for-people-with-disabilities-183057.

Hanson, R. (2016). *The age of em*. Oxford: Oxford University Press.

Krueger, J., & Osler, L. (2022). Communing with the Dead Online: Chatbots, grief, and Continuing Bonds. *Journal of Consciousness Studies*, *29*, 222–252.

Luck, M. (2009). The gamer's dilemma: An analysis of the arguments for the moral distinctions between virtual murder and virtual paedophilia. *Ethics and Information Technology*, *11*, 31–36.

Millar, J. (2014). ‘Proxy Prudence: Rethinking Models of Responsibility for Semi-Autonomous Robots’. Available at SSRN: https://ssrn.com/abstract=2442273.

Nozick, R. (2014). ‘Philosophical Explanations’, *Essays and Reviews: 1959–2002*. 187–196.

Patridge, S. (2011). The incorrigible social meaning of Video Game Imagery. *Ethics and Information Technology*, *13*, 303–312.

Powers, T. (2003). Real wrongs in virtual communities. *Ethics and Information Technology*, *5*, 191–198.

Rubel, A., Castro, C., & Pham, A. (2021). Introduction. *Algorithms and autonomy: The ethics of automated decision systems*. Cambridge: Cambridge University Press.

Sweeney, P. (forthcoming). *Social robots: A fictional dualism model*. London: Rowman and Littlefield.

Sweeney, P. (2022). Why indirect harms do not support social robot rights. *Minds and Machines, 32*, 735–749.

Thornton, R. G. (2010). Responsibility for the acts of others. *Proceedings (Baylor University. Medical Center)*, *23*(3), 313–315.

Walker, M. (2014). Uploading and personal identity. In R. Blackford, & D. Broderick (Eds.), *Intelligence unbound: The Future of Uploaded and Machine Ethics*. Wiley & Sons.

Wolfendale, J. (2007). My avatar, my self: Virtual harm and attachment. *Ethics and Information Technology*, *9*(2), 111–119.