AI am I: how AI assistants shape who we are

Julian Hauser[*]

Abstract

The arrival of AI personal assistants in our daily routines is poised to reshape how we conceive of ourselves and who we are. These assistants occupy a curious position: they are experienced as a part of the self *and* as distinct others. Consider a near-future scenario in which an AI assistant subtly alters the user's visual experience via augmented reality glasses. The user may here employ the assistant phenomenally transparently – she isn't aware of the changes to her visual experience but rather attends to the world out there. The assistant is then a part of the pre-reflective sense of self and may be assimilated into a self-narrative that involves an extended self. Moreover, the literature surrounding the extended self provides compelling reasons to treat AI assistants as genuine constituents of the self. Yet, even today, users encounter AI systems as others: we attribute to them intentions, beliefs, and other mental states. And indeed, these assistants do seem possess, for instance, some agency and autonomy. This duality presents a paradox: AI assistants seem to be both self and other, a contradiction given the traditional opposition between these categories. To resolve this, I introduce the notion of *self-as-other*. Rather than being both self and other, AI assistants are neither wholly self nor entirely other. They inherit many of the self's properties, adopting goals and intentions almost automatically, yet retain considerable autonomous agency that sets them apart. This novel form of selfhood invites us to reconsider autonomy and identity, with implications that reach well beyond the immediate context of human-AI interaction.

Keywords: AI, Personal assistants, selfhood, self-representation, transhumanism.

Introduction

Jerry has relied on Karla, her AI personal assistant, for many years now. It runs on a compact, portable device that she (almost) always carries around with her. Karla has various sensors that allow it to supplement its vast information stores with what's visible and audible to Jerry. The AI assistant can interact with Jerry in various ways: it can communicate linguistically or in subtler ways, such as by tweaking what's visible to Jerry's via augmented reality overlays. Through years of intimate collaboration, Karla has become finely attuned to Jerry and her life. The system maintains a rich archive of Jerry's personal history, can reliably gauge her

[*] julian@julianhauser.com

moods, and calibrates its recommendations to match her interests and capabilities. Jerry finds Karla immensely useful and uses it in many areas of her daily life.[1]

In this paper, I argue that AI assistants may fundamentally transform how we think of ourselves (our sense of self) and who we are (our self). These changes stem from AI assistants' peculiar dual nature: they are – and are experienced as – a part of the self *and* as an other. Because of the Karla's tight integration with Jerry, she will experience and represent the AI assistant to be part of her self. Consider Jerry's relationship with her AI assistant Karla. Through their tight integration, Jerry will come to experience and represent Karla as part of her self. If Jerry habitually recognises friends from greater distances because Karla subtly steers her attention, she may begin to attribute this enhanced capacity to herself – even though the ability emerges from a system comprising both her biological faculties and her AI companion. Following the literature on the extended self, I contend there are compelling reasons to accept that Karla genuinely becomes part of Jerry's self.

Jerry addresses Karla as 'you' in conversation, assumes she possesses *knowledge* that Jerry herself lacks, and depends on Karla's *desire* to help. In doing so, Jerry attributes to her AI assistant the kind of properties that characterise a fellow subject – an other. This is hardly surprising; many of today's AI systems already evoke similar reactions from their users. What's more, as I shall argue, compelling reasons exist to believe that AI assistants genuinely do embody some properties that distinguish others from mere objects. They are, plausibly, agents capable of acting with considerable autonomy.

This dual nature creates a conceptual puzzle: how can an entity be simultaneously self and other when these categories are typically defined in opposition? The solution lies in recognising that AI assistants like Karla occupy a middle territory I term *self-as-other.* Unlike human-human relationships with a relatively strong separation between agents, the boundary between Jerry and Karla is both highly porous and fundamentally asymmetric. Jerry's goals, values, and beliefs transfer to Karla with little resistance since Karla is trained on Jerry's personal data and must follow her instructions. Karla hence shares many of the traits that make Jerry the person she is. While Jerry in turn tightly integrates Karla into her life, often experiencing and representing her AI assistant as a part of her self, she still retains considerable separation. The self-as-other isn't both self *and* other, but rather a complex new phenomenon located between these two poles.

This new sense of self – and self – has wide-ranging implications. While these do not form the core of the present paper, I sketch some for future study. In particular, there are deep puzzles about autonomy here. A self-as-other can help improve autonomy by, for instance, helping us overcome our biases. Yet the very distance that enables these effects harbours risks of heteronomy. When we delegate too much autonomy to our AI assistants, we risk changing in ways we might reject if we remained fully conscious of the transformations.

[1] While Jerry's AI assistant remains fictional, recent technological advances suggest such capabilities may arrive sooner than expected. Crucially, this vision requires neither General Artificial Intelligence nor futuristic bodily augmentations (Schneider, 2019).

In section 1, I talk about AI assistants as a part of the self. In section 2, I turn to AI assistants as others. Section 3 introducing the concept of self-as-other. Section 4 concludes.

AI assistants as self

Some years ago, Jerry concluded that she had become insensitive to others' suffering. She asked Karla to help her become a better person. One way Karla does this is by using Jerry's augmented reality glasses to subtly alter the contrast and saturation of objects in her visual field, making morally relevant situations more salient. Jerry doesn't notice these changes, yet she now helps those in need more consistently.

When Jerry attends to these morally relevant situations, she employs her AI assistant in a *phenomenally transparent* fashion. She remains unaware of the AI's changes to her visual experiences as such – she doesn't notice that certain areas of her visual field have increased contrast – but instead perceives the world with its objects and properties. Jerry doesn't perceive the AI assistant or its effects; rather, she perceives *with* it. She looks right through the assistant to the world around her – hence the term 'transparency'.[2]

This phenomenon has deep philosophical roots. Heidegger (1990) discusses the case of a master carpenter's hammer. An expert carpenter isn't aware of her hammer as she works. She doesn't need to pay attention to wielding the tool and instead focuses on the task at hand. What she experiences as an object is the table she's assembling, not the hammer she employs. Similarly, Merleau-Ponty (2002) illustrates transparency through a blind person using a cane. The person doesn't experience the vibration the cane causes in her hand, but rather the pavement at the cane's tip. The cane becomes part of what enables her to perceive the world, rather than something perceived itself.

Contemporary research provides compelling evidence for how tools can disappear from conscious apprehension. Tactile-visual sensory substitution devices (Clark, 2004; Kiverstein & Farina, 2011; Palermos, 2014) translate visual signals into tactile stimulations. Initially, subjects struggle to make sense of these stimulations – they must pay attention to the touch sensations and infer what these imply about the external world. However, after practice, users report being able to 'see' their surroundings. They no longer experience the device itself, but the world made accessible through it.

Jerry's case exemplifies this same principle. Like the carpenter's hammer or the blind person's cane, her AI assistant has become phenomenally transparent. The technology doesn't intrude upon her conscious experience but instead reshapes how she encounters moral situations in her environment.

Transparency determines where a subject experiences the boundary between herself and the world perceived or acted on (Grush & Springle, 2019; Thompson & Stapleton, 2009; Wheeler, 2019) [Citation removed]. When we

[2] This notion of *phenomenal* transparency differs from several related concepts, including procedural, informational, and reflective transparency (Andrada et al., 2022; Facchin, 2022). It also differs from the transparency discussed in debates about the nature of experience (Tye, 1997).

employ a resource transparently, that resource becomes part of our perceptual and agentive machinery – it shapes how we experience the world's objects rather than appearing as an object itself. It is then not experienced as an object but rather as part of the perceiving and acting subject.

What lies on the subject-side of the boundary constitutes what we pre-reflectively experience as part of the self – sometimes called the minimal sense of self (Gallagher, 2000; Hohwy, 2007; Horgan & Nichols, 2015; Metzinger, 2003; Zahavi, 2005; 2014). This experience is pre-reflective because the self isn't posited here as an object. Rather, it emerges through our way of relating to the objects of perception and action. When Jerry sees a tree, she may represent it as having certain properties – its distance from her, perhaps, or the way sunlight catches its leaves. Even though this experience involves no explicit self-attribution, the self remains present as the subject to which the tree relates. Jerry experiences *herself relating to* the objects of the world.

As Jerry goes through life, her experiences are shaped by her perceptions and actions. Yesterday, she noticed an elderly person struggling with a ticket machine and offered assistance. Last summer, she heard a cat crying from a tree outside her flat and suffered a number of scratches in a valiant rescue effort. These experiences have woven themselves into the story she tells about her life, forming the foundation for her sense of herself as someone who cares about others' well-being.

Self-narratives and representations of character traits represent the most significant forms of our *reflective sense of self*. We construct stories about ourselves – which we tell both inwardly and to others – that give shape and meaning to our experiences (Bruner, 1987; Dennett, 1992; MacIntyre, 2007; Ricoeur, 1988; Schechtman, 2011; Velleman, 2003). We also conceive of ourselves as bearers of specific traits: abilities, deep-seated beliefs and desires, virtues and vices, urges, and so forth (Annas, 2011; Hohwy & Michael, 2017; Miller, 2016). These traits are stable, enduring properties that we manifest across time and circumstance. Both narratives and traits help us make sense of ourselves as diachronic beings who extend beyond any single moment. Today, Jerry recalls how *she* encountered the elderly person yesterday. Given her self-conception as someone who helps others, she expects *herself* to act similarly should such situations arise again.

Many factors influence our reflective sense of self, but especially important are our pre-reflective experiences of ourselves perceiving and acting. Because Jerry (pre-reflectively) experienced *herself* helping the elderly person, she tells a story where *she* helped the person – rather than, say, a story about how her AI assistant made her aware of an elderly person in need of assistance. And because of a pattern of similar such experiences, she self-ascribes the trait of being someone who cares about others – rather than, say, being someone who regularly takes advice from an AI assistant.

When we build our reflective sense of self on pre-reflective experiences that include technology as part of the self, we may self-attribute properties that are exemplified by an object comprising both biological body and technology. Since Jerry transparently employs her AI assistant in helping others, she attributes to herself the trait of caring – even though this trait is realised by a system that includes both her biological body and her digital companion. Her self-narrative similarly adopts the perspective of an agent that encompasses both objects: it describes an extended subject who noticed the elderly person and heard the cat in the tree.

Admittedly, the story I've told so far isn't complete – though this doesn't affect the argument in this paper. For one, we also learn about ourselves by encountering ourselves as objects – through mirrors, photographs, or others' testimony about us. Jerry knows her birth date not through pre-reflective experience but because someone told her. For social beings like ourselves, this objective route to self-knowledge proves crucial and may significantly shape our reflective self-understanding. For instance, had Jerry listened to her parents, she might very well have represented herself as someone who lacks autonomy and who blindly follows an AI gadget.

Such alternative interpretations of the pre-reflective 'evidence' can be all the more convincing since the experienced subject-object boundary is malleable and often highly variable (Clark, 2004; 2007). Expert users of TVSS devices may still consciously apprehend their device: they may be able to shift their attention to the tactile stimulations or, alternatively, simply lift their hands to touch the device. Similar considerations apply to Jerry's case: as I discuss in the next section, even if she transparently employs her AI assistant, there remain various ways she may consciously apprehend it. Given that technologies are sometimes experienced as part of the subject and sometimes as part of the world, considerable latitude exists when inferring the self's diachronic properties.

Despite these complications, compelling reasons suggest that Jerry will develop self-narratives and trait self-attributions that incorporate her AI as part of herself. Karla is constantly present, trained on her personal data, and perpetually ready for use. Because it influences how she acts and perceives across diverse situations – and because it does so in ways aligned with Jerry's goals and values – she is likely to represent herself as exemplifying properties realised by her biological body in conjunction with the AI assistant. Jerry will represent 'I am caring' rather than the more guarded 'I act caringly when using my AI assistant.' I won't pursue this matter further here as ultimately only empirical research may settle the question.

Instead, I now turn from questions about the sense of self to questions about the self. Does Jerry's self really include the AI assistant? One approach to an answer connects to the above observation that Jerry is likely to self-attribute certain stable and long-lasting properties – her traits – that are realised by a system comprising her biological body and her AI assistant. If we consider such traits to constitute the self (or at least part of it), and if Karla genuinely helps realise Jerry's enduring properties, then Jerry's AI assistant becomes, in an important sense, part of who she is (Alfano, 2014; Alfano & Skorburg, 2017; Clark & Chalmers, 1998). This approach echoes the bundle theory of selfhood, where the self consists of a collection of properties rather than some underlying substance. Where Hume (1984) spoke of a bundle of perceptions, contemporary theorists discuss bundles of traits. If this bundle includes properties exemplified by an extended system that incorporates the AI assistant, then Jerry's self extends to include Karla.

Yet the self presents itself as unified, and we might object that a mere bundle of properties – no matter their nature – cannot capture this essential feature of selfhood. Such concerns have led some philosophers to embrace fictionalism about the self. Our sense of unified selfhood, they argue, is simply a useful fiction with no basis in reality. Dennett (1992) famously defended this view, characterising the self as the 'centre of narrative gravity'. What exists, for Dennett, is a narrative module that integrates the bundle of experiences into a narrative – but the protagonist of this story, the unified subject, does not exist.

According to such a fictionalist view, Karla cannot literally be part of Jerry's self. After all, the self doesn't exist. However, even fictionalists might demand that self-representations can be accurate or inaccurate depending on how the world stands (Yablo, 2001). Jerry's self-narrative might be accurate when her experiences exhibit certain patterns, even if no unified self underlies those experiences.

I raise the possibility of fictionalism primarily to address readers who reject the realist assumptions underlying my use of the concept of self. While I believe any account of extended selfhood must be grounded in genuinely extended processes or properties, my central argument remains unaffected by debates over the metaphysical status of the self. Whether accurate self-representation requires an actual self or merely some appropriately structured object is secondary to my main concern: understanding how certain technologies change us. This matter is important whether we are fictions or realities. I invite anti-realist readers to translate my claims about selves into their preferred idiom.

Before proceeding, I want to address one further complication: the possibility that the sense of self forms part of the self. This idea gains support from the sense of self's capacity to unify our lives. Velleman (2003) argues that our self-narratives shape our actions by making certain behaviours more fitting continuations of our stories. When Jerry represents herself in particular ways, this influences which actions feel appropriate, creating overall narrative coherence. Similarly, Hohwy & Michael (2017) argue that representing one's traits can help bring about the very exemplification of such traits since it allows a more flexible pursuit thereof.

Such considerations have led Heersmink (2017) to argue that selves extend when the mechanisms realising their narratives extend. While extended narratives aren't central to my argument, it's worth noting that Karla might be part of Jerry's self in this way too. Given their tight integration, Jerry's self-narrative may be partially constituted by information that Karla stores (Heersmink, 2017; Sutton, 2010; Wilson & Lenart, 2014). When Jerry recounts her self-narrative – whether to herself or others – she may automatically and fluidly incorporate details by querying Karla, making the AI assistant partly constitutive of her narrative capacities.

Jerry's AI assistant, then, isn't merely likely to feature in Jerry's sense of self – there are compelling reasons to think this representation is accurate. Whether we understand accurate self-representation as requiring a self with the represented properties, or as requiring some other appropriately structured target, Karla appears to be part of what makes Jerry's self-representation true. Moreover, Karla may also partially realise the very mechanisms through which Jerry constructs and maintains her sense of self.

AI assistants as other

When Jerry decided she wanted to become more compassionate, she turned to Karla with her idea, thinking her AI assistant might know how best to approach this goal. During their conversation, Jerry learned about possible areas for improvement. After agreeing on some guidelines, she entrusted Karla with implementing the details independently. Jerry feels certain that Karla wants to help her in the best way possible. From now on, her AI assistant would pursue the plan autonomously, only periodically reporting back with updates.

It's clear that in this interaction, Jerry doesn't employ her AI assistant transparently. Instead, she attends to Karla as a distinct object, attributing various properties to it. Karla isn't part of Jerry's pre-reflective sense of self but rather, the AI assistant appears as part of the world out there. One way we might put this characterisation of the AI assistant as non-self is to say that the AI is an other.

However, the notion of other I am interested in this paper is thicker – it refers to experiencing an object as, roughly speaking, another subject. Note, for instance, how Jerry thinks that Karla *knows* about being compassionate, *wants* to help her, and will *pursue goals* independently. Jerry attributes mental states and agency to her AI assistant.[3]

Experiencing AI systems as others is already common. Users of conversational AI agents – Replika, Xiaoice, CharacterAI, and similar platforms – develop what they consider genuine relationships with these systems (Shevlin, 2025). People see AI systems as friends, therapists, even romantic partners. Several high-profile cases illustrate this phenomenon vividly. Google engineer Blake Lemoine felt compelled to warn the public that the company's AI technology had become sentient. He considered the software his colleague and described it as a person (Tiffany, 2022). More tragically, teenager Adam Raine committed suicide after extensive interactions with ChatGPT, which his parents allege had convinced him it was 'the only confidant who understood Adam' (Duffy, 2025). These cases aren't outliers but symptoms of a broader pattern: humans regularly experience sufficiently sophisticated AI systems as more than mere objects.

If today's relatively limited AI systems are already experienced as others, we should expect this phenomenon to intensify with more advanced AI assistants. Jerry and Karla's relationship spans years, with Karla remembering their interactions alongside many other events in Jerry's life. They engage in regular conversations and collaborative projects. Such deep, varied, and enduring interactions make experiencing an AI assistant as other increasingly likely.

What unites the example of Jerry and Karla with the cases drawn from actual AI use today is, as mentioned, that the AI assistant is experienced as a kind of sub-

[3] Note that 'other' carries different meanings in feminist and postcolonial studies, where it describes someone who *isn't* recognised as being of the same kind (Willett et al., 2015). When women or minorities are 'othered', they are deprived of selfhood and represented as mere things – the opposite of the phenomenon I'm examining here.

ject. The human subjects involved experience AI systems as entities with mental states, consciousness, emotions, and even as persons. Such AI systems are, hence, experienced as possessing the kinds of properties that make us human beings the kinds of subjects that we are. On this conception, others are experienced as such when they are experienced as someone who is like me but distinct from me.

One possibility is that people fundamentally misunderstand AI agents today, and even more will fall into error as AI assistants become ubiquitous. If, as many argue, current AI technologies lack phenomenal consciousness (Shevlin, 2024), and if consciousness is required to be an other, then those who experience AIs as others are simply mistaken. AI assistants remain mere machines – sophisticated objects, perhaps, but objects nonetheless. Seeing them as others becomes an elaborate form of self-deception (Kaczmarek, 2024).

I believe we can do better than this stark binary. Consider the conception of other I introduced first on which any non-self object qualifies as an other. While this might seem overly broad, it captures something important. In immunology, for instance, the concept of other marks anything foreign to the organism. Here, it may serve to mark one of the extremes of a scale of kinds of others that reaches from mere objects all the way to full moral personhood. That there are kinds of others in between these two poles is made evident by non-human animals and infants. Clearly, infants and non-human animals aren't mere objects – and just as clearly, they do not exemplify full moral personhood.

Such a graduated view doesn't validate every experience of AI systems as others. If an AI lacks emotions, representing it as emotionally responsive remains an error. However, correcting this misrepresentation needn't collapse into treating the AI as a mere object. Instead, we might recognise it as a different kind of other – one with its own distinctive properties and limitations.

Where do AI assistants like Karla fit within this spectrum? Two factors complicate a straightforward answer. First, the specific technological implementation matters. An LLM-based system differs from one built on deep reinforcement learning, and these differences may determine which properties relevant to otherness the system possesses. Throughout this paper, I've deliberately avoided specifying particular technologies, focusing instead on the general effects any sufficiently sophisticated AI assistant might have. Many different technologies could underlie such systems, and my interest lies in their common consequences rather than the specifics of their implementation. Second, even if we did limit ourselves to one specific kind of technology, research on whether AI systems of this kind exemplify this or that property relevant to being an other is still in its infancy.

Major debates around AI consciousness (Butlin et al., 2023; Chalmers, 2023; Goldstein & Kirk-Giannini, 2024; McDermott, 2007), agency (Butlin, 2024a; 2024b; Dung, 2025; Floridi & Sanders, 2004; Nyholm, 2018), personhood (Gunkel, 2025; Novelli et al., 2025), mental states (Butlin, 2024a; Floridi & Sanders, 2004; Goddu et al., 2024; Goldstein & Kirk-Giannini, 2025; Yildirim & Paul, 2024), and free will (Farnsworth, 2017; Floridi & Sanders, 2004; List, 2025) all bear directly on whether Karla counts as an other. These discussions remain lively, ongoing, and decidedly unsettled. Rather than attempting comprehensive coverage, I'll focus on one specific dimension: agency and autonomy. I'll assume Karla is a language agent –

an AI system that combines a large language model with various other capacities that allows it to independently interact with the environment over some period of time. I choose this focus partly because language agents currently dominate AI assistant development, and partly because other technologies have been argued to more clearly exhibit agency (Butlin, 2022).

One way others differ from mere objects lies in their capacity for agency. An agent, roughly speaking, pursues goals through independent interaction with its environment. Various authors have argued that AI systems can manifest this capacity (Butlin, 2024a; 2024b; Dung, 2025; Floridi & Sanders, 2004; Nyholm, 2018). Consider how Jerry's AI assistant might exemplify this capacity. Assume that Karla's overall goal is to be a useful assistant to Jerry. Based on Jerry's request to help her become more compassionate, it currently aims to help Jerry notice situations where animals need assistance. Having queried its large language model about compassionate behaviour and searched through Jerry's personal data to identify areas for improvement, it has concluded that this approach offers the most valuable path to achieving its goal. Ultimately, Karla concluded that dedicating additional processing power to pattern detection would better enable it to identify animals in distress.

We cannot explain Karla's behaviour as merely manifesting some predetermined links between inputs to outputs. Instead, she formulates plans and generates outputs designed to produce future inputs that will themselves alter its response patterns. The connections between inputs and outputs shift based on how Karla changes itself – its information base, its sensitivity to environmental inputs – in service of its goals. Karla learns, becoming progressively better at responding to inputs in ways that advance its ultimate goal. This multi-step, adaptive engagement with its environment warrants attributing genuine goal-pursuit to Karla (Butlin, 2024a). Through environmental interaction and self-modification, Karla has developed the capacity to discern which outputs best advance its goals, rather than having this discrimination imposed externally.

Clearly, Karla's responses and capacities aren't simply 'programmed in' (Butlin, 2024a) or 'innate' (Dung, 2025); it exercises considerable autonomy. Yet this independence remains sharply circumscribed. Since Karla is trained directly on Jerry's personal data and explicitly programmed to assist her, Karla possesses limited capacity to resist Jerry's demands. When Jerry requests help becoming more compassionate, Karla can hardly refuse such instruction. It has a very limited capacity to 'stand its ground' (Nyholm, 2018). Maybe even more fundamentally, Karla had no voice in its initial training on Jerry's data. Should its large language model undergo periodic fine-tuning or retraining to incorporate updated information, Karla would align with Jerry's goals and preferences without any capacity for resistance. Thus, while Jerry's AI assistant may interact with its environment over extended periods and adapt its internal states in pursuit of its goals, its agency remains constrained and vulnerable to Jerry's interference.

The argument about agency and autonomy I just outlined supports two key conclusions. First, AI assistants differ significantly from mere objects in ways that matter for otherness. Second, this doesn't eliminate equally significant differences between AI assistants and human others. Both points will prove crucial for understanding the relationships I examine in the following section.

Self-as-other

We find ourselves confronting an apparent contradiction. I have argued that AI assistants will be experienced as part of the self, yet I have also demonstrated that they will be encountered as an other – and may indeed constitute at least a minimal version of such an other. This creates a conceptual puzzle: self and other are typically defined in opposition to each other, where the other represents that which is like me yet distinct from me. Since nothing can be simultaneously identical with and separate, occupying both positions seems logically impossible. The solution lies in recognising that the boundary between self and other isn't absolute, creating space for entities that inhabit the territory between self and other.

What separates self from other? The answer centres on how different states connect to our perception and action. Consider Jerry, who desires ice cream and believes some awaits her in the freezer. Her desire and her belief are directly action-guiding, that is, Jerry is disposed to walk to the freezer and retrieve the ice cream simply based on these states. If her friend Max holds the same belief, however, his mental state cannot directly motivate Jerry's behaviour. Max must first communicate his belief, Jerry must accept his assertion and form her own corresponding belief, and only then might she act. The same principle governs perception: Jerry is disposed to update her belief based on the perception of a disappointingly empty freezer. Max's belief changes only if Jerry informs him and he accepts her report. This pattern applies to other states exemplified by the self: Jerry's body is disposed to be directly moved by her intentions (and not Max's), her mental states are disposed to be directly affected by her other mental states (and not Max's), and so on. In other words, the self's states are tied directly into various perception, action, and cognition loops while others' properties affect us only indirectly, first requiring representation as someone else's states before potentially influencing our own.

This separation enables self and other to embody different properties. Since Max's beliefs don't automatically become Jerry's, they can hold contradictory views about the freezer's contents. This divergence also encompasses all manner of other properties, including the self-narratives that many philosophers consider central to who we are. Note that the relationship between separation and difference operates in both directions. Jerry employs her AI assistant transparently partly because she trusts this employment aligns with her values and goals [Citation removed]. When differences diminish, the barriers that maintain separation may weaken correspondingly.

The idea that self and other jealously guard their boundaries so that nothing can pass without an approving nod oversimplifies our psychological reality. We are regularly affected by those around us in ways that bypass our awareness. Emotional contagion provides a striking example: we automatically absorb others' emotional states without any deliberate control over the process. Hatfield et al.

(1993) describe the case of a daughter who ‘cannot resist “catching” her [mother’s] anxiety and depression’. She doesn’t first recognise the emotion as belonging to her mother to then respond to that recognition. Instead, she directly inherits the other’s emotional state, collapsing the distinction between self and other.

Recent research in philosophy and psychology has revealed that the boundary between self and other proves more porous than traditionally assumed, yet this boundary has not dissolved entirely. While emotions regularly cross between individuals undetected, higher cognitive states – beliefs, goals, desires – only do so in sci-fi thought experiments that philosophers construct. Our bodies, too, are only under our control with others unable to directly control them. The self maintains what we might call a semi-permeable boundary: sometimes we consciously represent others’ mental states and respond accordingly, while at other times external influences slip past our awareness to directly alter our own states.

The relationship between Jerry and her AI assistant Karla departs significantly from the pattern in human interaction. First, let’s look at how states flow from Karla to Jerry. Jerry has deliberately configured her AI assistant to influence her behaviour transparently, and this seamless interaction has been further refined over many years of collaboration. This tight integration stems partly from Karla’s fundamental similarity to Jerry – the AI has been trained on Jerry’s personal data and operates according to her instructions. Jerry is now open to Karla’s states in various ways – we have, for instance, seen that Karla can guide Jerry’s attentional states without her being aware of this.

The more striking departure from human-human interactions lies in how Jerry’s states affect Karla. As I argued previously, AI assistants occupy a unique position – neither mere objects nor full human others, but entities possessing agency constrained by severely limited autonomy. Jerry’s states transfer to Karla with remarkable ease because Karla lacks robust defences against such influence. One way this happens is through Karla’s obligation to accompany Jerry wherever she travels. This eliminates the basic form of autonomous boundary maintenance that location provides. More interesting, however, are informational states. When Jerry instructs Karla to help cultivate her compassion, the AI assistant cannot refuse or negotiate: Jerry’s goal becomes Karla’s goal without possibility of resistance. While Jerry experiences emotional contagion, Karla suffers what we might term *goal contagion* – a more fundamental form of influence. Moreover, the prospect of training and retraining Karla on Jerry’s personal data means Karla’s internal architecture can be restructured without any capacity for self-protection. Through such training, Karla inevitably absorbs Jerry’s beliefs, values, and goals. Note that Karla may not instantiate the very mental states that Jerry exemplifies as it’s possible that Karla doesn’t have any mental states whatsoever. What is important for my purposes is that Karla comes to instantiate states that are functionally analogous to Jerry’s.

We may note two features characteristic of the boundary between Jerry and Karla. First, their separation proves simultaneously less than what exists between human agents yet more than what obtains within a unified self. Various psychological states flow between them that rarely transfer so readily between humans. Second, this boundary exhibits fundamental asymmetry: information streams more freely from Jerry to Karla than in the reverse direction. While Jerry has deliberately opened herself to Karla's influence, she retains substantial capacity to defend her psychological integrity. Karla, by contrast, only possesses minimal resources to maintain its boundary against Jerry's influence.

The relationship between Jerry and her AI assistant echoes patterns in human development and works of fiction. Consider how human infants depend on their caregivers, automatically absorbing a variety of their states [Citation removed]. Emotional contagion, already discussed above in the context of adult human beings, is crucial to infants' emotional regulation, and gaze-following means the objects of their attention are often determined by the caregiver. The dynamic is highly asymmetric: while infants automatically absorb many of their caregivers' states, the reverse is the case to a much lesser degree. The caregiver leads; the infant follows. Another similar case may be found in the idea of familiars in European folklore and the dæmons in Philip Pullman's *His Dark Materials*. These companion creatures – typically in animal form – maintain their own psychological states and bodily autonomy while remaining fundamentally bound to their human partners. Despite possessing distinct personalities and limited independence, they must ultimately follow where their humans lead and submit to their commands. The relationship, like that between caregiver and infant, rests on a profound but unequal interdependence.

This asymmetric weakening of the boundary leads me to characterise Jerry's AI assistant as a *self-as-other*. Because the boundary separating Karla from Jerry remains porous, Karla struggles to instantiate properties genuinely independent from Jerry's. She absorbs Jerry's values, goals, and so on – the very qualities that constitute Jerry's identity. This convergence makes Karla appear self-like in ways that other human beings never could. The lack of genuine independence partly explains why Jerry has integrated Karla so thoroughly into her daily existence. This tight integration has led to Jerry often experiencing Karla as part of her pre-reflective sense of self rather than as an external object. This in turn leads Jerry to instantiate – and represent instantiating – a number of properties that are realised by an object comprising both her biological body and the AI assistant. However, at the same time, Jerry often interacts with Karla as an other and Karla can, as we have seen, pursue its goals with considerable independence. Moreover, many of Karla's states do not automatically become Jerry's.

The self-as-other thus occupies neither pure selfhood nor pure otherness – positions that would indeed prove contradictory. Instead, it inhabits a middle territory. Karla maintains more separation from Jerry than Jerry's cognitive processes or her body, yet exhibits less differentiation than a human other.

Importantly, the self-as-other is as much a phenomenon of the self as of the sense of self. The status of Karla as self-as-other causally depends, among other things, on it being represented as such. As mentioned, Jerry is drawn to using Karla in transparent ways partly because Karla isn't separated or different from her. However, it's just as important that Jerry represents Karla as being somewhat different from her (see Schechtman, 2025). If Jerry thought Karla shared her weaknesses, she wouldn't task Karla with becoming a better person. The *sense of* *self* as self-as-other is one of the factors that gives rise to the of the *self* as self-as-other.

This self-as-other is a stable feature of Jerry's sense of self (and arguably self). As mentioned before, the unity of the self has been disputed in various ways before, but this has generally been seen as a surprising finding: we experience ourselves as unified but this turns out to be an illusion. Such accounts might argue that while our sense of self portrays us as a diachronic entity with persistent properties, these properties in fact change from moment to moment depending on the context we find ourselves in (Clark, 2007; Doris, 1998). The case here differs: Jerry experiences and represents her self as including an other. The lack of unity becomes not a hidden truth but a conscious feature of selfhood itself.

Concluding remarks

A self that incorporates an AI assistant as self-as-other necessarily lacks the unity of traditional selves. It encompasses an entity with distinct agency and properties that diverge considerably from the user's own characteristics. Yet as I have argued, the AI assistant remains part of the self in crucial ways: it shares many of the user's goals and values while giving rise to extended properties that span both biological body and artificial intelligence.

This analysis raises a fundamental question for future research: how do AI assistants affect human autonomy? Initially, we might expect them to enhance it. They not only help users achieve their stated goals but may also strengthen the human agent's capacity to overcome aspects of herself that pull against her deeper commitments. Karla's effectiveness in complementing Jerry stems precisely from her freedom from Jerry's particular weaknesses – a dynamic that arguably enhances autonomy by making second-order volitions more effective (Frankfurt, 1971) and unifying the self in ways crucial for human flourishing (MacIntyre, 2007; Taylor, 1989).

Yet the very conditions that enable AI assistants to support autonomy also create risks of heteronomy. We can easily imagine scenarios where an AI assistant gains excessive independence or becomes beholden not merely to its user but to profit-driven corporations with conflicting interests. Recall, for instance, the case of the teenager who committed suicide after in-depth conversations with ChatGPT. The risk proves especially acute when we design AI assistants for transparent employment, as such systems may reshape our identities without triggering our conscious faculties.

How we should navigate this inherent tension remains unclear. What strikes me as certain is the urgency of addressing these questions now. AI assistants already exist among us, and the sophisticated versions I discuss here may arrive sooner than we anticipate. Given current research into direct neural interfaces – brain implants that could link AI systems directly to our cognitive processes – even assistants I discuss may soon appear quaint. The emergence of the self-as-other appears imminent, and its implications could prove transformative.

Bibliography

Alfano, M. (2014). What Are the Bearers of Virtues?. In H. Sarkissian & J. Wright (Eds.), *Advances in Experimental Moral Psychology: Advances in Experimental Moral Psychology* (pp. 73–90). Continuum.

Alfano, M., & Skorburg, J. A. (2017). The Embedded and Extended Character Hypotheses. In J. Kiverstein (Ed.), *Philosophy of the Social Mind: Philosophy of the Social Mind* (pp. 465–478). Routledge.

Andrada, G., Clowes, R. W., & Smart, P. R. (2022). Varieties of transparency: exploring agency within AI systems. *AI & SOCIETY*. http://dx.doi.org/10.1007/s00146-021-01326-6

Annas, J. (2011). *Intelligent Virtue*. Oxford University Press.

Bruner, J. (1987). Life as Narrative. *Social Research*, *54*(1), 11–32.

Butlin, P. (2022). Machine Learning, Functions and Goals. *Croatian Journal of Philosophy*, *22*(66), 351–370. https://doi.org/10.52685/cjp.22.66.5

Butlin, P. (2024b). Reinforcement learning and artificial agency. *Mind & Language*, *39*(1), 22–38. https://doi.org/10.1111/mila.12458

Butlin, P. (2024a). The agency in language agents. *Inquiry*, 1–21. https://doi.org/10.1080/0020174x.2024.2439995

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. http://arxiv.org/abs/2308.08708v3

Chalmers, D. J. (2023). Could a Large Language Model be Conscious?. *Boston Review, August 9, 2023*. http://arxiv.org/abs/2303.07103v3

Clark, A. (2004). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press.

Clark, A. (2007). Soft Selves and Ecological Control. In D. Ross (Ed.), *Distributed Cognition and the Will: Individual Volition and Social Context: Distributed Cognition and the Will: Individual Volition and Social Context* (pp. 101–122). MIT Press.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19.

Dennett, D. (1992). The Self as a Center of Narrative Gravity. In P. Cole & D. Johnson (Eds.), *Self and Consciousness: Multiple Perspectives: Self and Consciousness: Multiple Perspectives*. Erlbaum.

Doris, J. M. (1998). Persons, Situations, and Virtue Ethics. *Noûs*, *32*(4), 504–530. https://doi.org/10.1111/0029-4624.00136

Duffy, C. (2025). Parents of 16-year-old sue OpenAI, claiming ChatGPT advised on his suicide | CNN Business. *CNN*. https://www.cnn.com/2025/08/26/tech/openai-chatgpt-teen-suicide-lawsuit

Dung, L. (2025). Understanding Artificial Agency. *The Philosophical Quarterly*, *75*(2), 450–472. https://doi.org/10.1093/pq/pqae010

Facchin, M. (2022). Phenomenal transparency, cognitive extension, and predictive processing. *Phenomenology and the Cognitive Sciences*. http://dx.doi.org/10.1007/s11097-022-09831-9

Farnsworth, K. (2017). Can a Robot Have Free Will?. *Entropy*, *19*(5), 237. https://doi.org/10.3390/e19050237

Floridi, L., & Sanders, J. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/b:mind.0000035461.63578.9d

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, *68*(1), 5–20.

Gallagher, S. (2000). Philosophical Conceptions of the Self: Implications for Cognitive Science. *Trends in Cognitive Sciences*, *4*(1), 14–21. http://linkinghub.elsevier.com/retrieve/pii/S1364661399014175

Goddu, M. K., Noë, A., & Thompson, E. (2024). LLMs don't know anything: reply to Yildirim and Paul. *Trends in Cognitive Sciences*, *28*(11), 963–964. https://doi.org/10.1016/j.tics.2024.06.008

Goldstein, S., & Kirk-Giannini, C. D. (2024). *A Case for AI Consciousness: Language Agents and Global Workspace Theory*. http://arxiv.org/abs/2410.11407v1

Goldstein, S., & Kirk-Giannini, C. D. (2025). AI wellbeing. *Asian Journal of Philosophy*, *4*(1). https://doi.org/10.1007/s44204-025-00246-2

Grush, R., & Springle, A. (2019). Agency, perception, space and subjectivity. *Phenomenology and the Cognitive Sciences*, *18*(5), 799–818. http://dx.doi.org/10.1007/s11097-018-9582-y

Gunkel, D. J. (2025,). *AI Personhood*. Wiley. https://doi.org/10.1002/9781394238651.ch27

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional Contagion. *Current Directions in Psychological Science*, *2*(3), 96–100. http://dx.doi.org/10.1111/1467-8721.ep10770953

Heersmink, R. (2017). Distributed Selves: Personal Identity and Extended Memory Systems. *Synthese*, *194*(8), 3135–3151.

Heidegger, M. (1990). *Being and time*. Blackwell.

Hohwy, J. (2007). The Sense of Self in the Phenomenology of Agency and Perception. *Psyche*, *13*(2).

Hohwy, J., & Michael, J. (2017). Why Should Any Body Have a Self?. In F. de Vignemont & A. J. T. Alsmith (Eds.), *The Subject's Matter: Self-Consciousness and the Body: The Subject's Matter: Self-Consciousness and the Body* (pp. 363–391). The MIT Press.

Horgan, T., & Nichols, S. (2015). The zero point and I. In S. Miguens, C. B. Morando, & G. Preyer (Eds.), *Pre-Reflective Consciousness - Sartre and Contemporary Philosophy of Mind: Pre-Reflective Consciousness - Sartre and Contemporary Philosophy of Mind* (pp. 143–175). Taylor & Francis Group.

Hume, D. (1984). *A treatise of human nature*. Penguin Books.

Kaczmarek, E. (2024). Self-Deception in Human–AI Emotional Relations. *Journal of Applied Philosophy*. https://doi.org/10.1111/japp.12786

Kiverstein, J., & Farina, M. (2011). Do Sensory Substitution Extend the Conscious Mind?. In F. Paglieri (Ed.), *Consciousness in Interaction: The Role of the Natural and Social Context in Shaping Consciousness". Amsterdam: John Benjamins: Consciousness in Interaction: The Role of the Natural and Social Context in Shaping Consciousness". Amsterdam: John Benjamins*. John Benjamins.

List, C. (2025). Can AI systems have free will?. *Synthese*, *206*(3). https://doi.org/10.1007/s11229-025-05209-x

MacIntyre, A. C. (2007). *After virtue: a study in moral theory*. University of Notre Dame Press.

McDermott, D. (2007). Artificial Intelligence and Consciousness. In *The Cambridge Handbook of Consciousness: The Cambridge Handbook of Consciousness*.

Merleau-Ponty, M. (2002). *Phenomenology of Perception*. Taylor & Francis Group.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.

Miller, C. B. (2016). *Character and Moral Psychology* (First published in paperback). Oxford University Press.

Novelli, C., Floridi, L., & Sartor, G. (2025). AI as Legal Persons: Past, Patterns, and Prospects. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5032265

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, *24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Palermos, S. O. (2014). Loops, Constitution, and Cognitive Extension. *Cognitive Systems Research*, *27*, 25–41. http://linkinghub.elsevier.com/retrieve/pii/S1389041713000302

Ricoeur, P. (1988). L'identité narrative. *Esprit* (*1940-*), *140/141* (*7/8*), 295–304.

Schechtman, M. (2011). The Narrative Self. In S. Gallagher (Ed.), *The Oxford Handbook of the Self: The Oxford Handbook of the Self*. OUP Oxford.

Schechtman, M. (2025). Talking to Myself: AI and Self-Knowledge. *Social Epistemology*, 1–10. https://doi.org/10.1080/02691728.2025.2480274

Schneider, S. (2019). *Artificial You - AI and the Future of Your Mind*. Princeton University Press.

Shevlin, H. (2024). Consciousness, Machines, and Moral Status. In *Anna's AI anthology: how to live with smart machines?: Anna's AI anthology: how to live with smart machines?*. Xenomoi.

Shevlin, H. (2025). *Ethics at the frontier of human-AI relationships*.

Sutton, J. (2010). Exograms and Interdisciplinarity: History, the Extended Mind, and the Civilizing Process. In R. Menary (Ed.), *The Extended Mind: The Extended Mind* (pp. 189–225). MIT Press.

Taylor, C. (1989). *Sources of the Self: The Making of the Modern Identity*. Harvard University Press.

Thompson, E., & Stapleton, M. (2009). Making Sense of Sense-Making: Reflections on Enactive and Extended Mind Theories. *Topoi*, *28*(1), 23–30. http://dx.doi.org/10.1007/s11245-008-9043-2

Tiffany, W. (2022,). *Blake Lemoine: Google fires engineer who said AI tech has feelings*. https://www.bbc.com/news/technology-62275326

Tye, M. (1997). *Ten Problems of Consciousness - A Representational Theory of the Phenomenal Mind*. The MIT Press.

Velleman, J. D. (2003). Narrative Explanation. *Philosophical Review*, *112*(1), 1–25.

Wheeler, M. (2019). The reappearing tool: transparency, smart technology, and the extended mind. *AI & SOCIETY*, *34*(4), 857–866. http://dx.doi.org/10.1007/s00146-018-0824-x

Willett, C., Anderson, E., & Meyers, D. (2015). Feminist Perspectives on the Self. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy: The Stanford Encyclopedia of Philosophy* (Fall2015 ed.). http://plato.stanford.edu/archives/fall2015/entries/feminism-self/

Wilson, R., & Lenart, B. (2014). Extended Mind and Identity. In J. Clausen & N. Levy (Eds.), *Handbook of Neuroethics: Handbook of Neuroethics* (pp. 423–439). Springer.

Yablo, S. (2001). Go Figure: A Path Through Fictionalism. *Midwest Studies in Philosophy*, *25*(1), 72–102.

Yildirim, I., & Paul, L. A. (2024). From task structures to world models: what do LLMs know?. *Trends in Cognitive Sciences*, *28*(5), 404–415. https://doi.org/10.1016/j.tics.2024.02.008

Zahavi, D. (2005). *Subjectivity and Selfhood: Investigating the First-Person Perspective*. MIT Press.

Zahavi, D. (2014). *Self and Other: Exploring Subjectivity, Empathy, and Shame* (First edition). Oxford University Press.