Forthcoming in *Topoi* special issue: “Affectivity and Technology: Philosophical Explorations”, eds. Marco Fasoli, Richard Heersmink, and Giulia Piredda.

Real Feeling and Fictional Time in Human-AI Interactions

Joel Krueger & Tom Roberts
University of Exeter

Abstract

As technology improves, artificial systems are increasingly able to behave in human-like ways: holding a conversation; providing information, advice, and support; or taking on the role of therapist, teacher, or counsellor. This enhanced behavioural complexity, we argue, encourages deeper forms of affective engagement on the part of the human user, with the artificial agent helping to stabilise, subdue, prolong, or intensify a person's emotional condition. Here, we defend a fictionalist account of human/AI interaction, according to which these encounters involve an elaborate practise of imaginative pretence: a make-believe in which the artificial agent is attributed a life of its own. We attend, specifically, to the temporal characteristics of these fictions, and to what we imagine artificial agents are doing when we are not looking at them.

Introduction

A range of modern technologies allow users to engage with artificial systems in ways that mimic elements of ordinary human interaction.[1] Some systems provide rich epistemic benefits (Alvarado 2023). They give the appearance of a knowledgeable interlocutor who can assist with directions, dates, opening times, and forgotten bits of

[1] We are very grateful for the feedback from two anonymous reviewers. Their comments helped us improve the paper. We’re also very grateful for Marco Facchin’s careful feedback on an earlier draft of this essay. Marco provided many helpful references, prompts further development, and helped us clarify several issues we hope to address in future work. Finally, we’d like to acknowledge the excellent comments and questions on this material from the participants of the 14 February 2024 Philosophy Colloquium at the University of Exeter, organised by Tyler Brunet. We’re fortunate to have such thoughtful colleagues.

trivia. Think of Jeeves from the early internet search engine, for instance, or more recent digital companions like Apple’s Siri and Amazon’s Alexa. Other interactive technologies fabricate more specific responsive personas such as a virtual customer service operative, travel agent, or bank teller. Advances in conversational AI, moreover, have enabled artificial systems of even greater sophistication such as virtual wellness coaches and CBT specialists, language tutors, or chatbots designed to aid the grieving process by replicating a loved one's patterns of speech (Buben 2015; Krueger & Osler 2022).[2] When a person engages with an artificial agent of this latter kind, the encounter is often charged with affective significance: the human user is an emotional creature with cares and concerns, fears, hopes, pains, and regrets, and the interactive technology plays a fundamental role in structuring and regulating these feelings. When things go well, these technologies can be used to confront and resolve negative emotions; improve a state of anxiety or depression; vent and undergo catharsis; or add stability and a sense of connection to a person's daily affective condition.[3]

There is an increasing amount of work on the affective character of these human-AI interactions (see Cavallo et al. 2018 for an overview). Much of this work focuses on emotional connections with service and social robots (e.g., Fussi 2023; Hung et al. 2019; Kerruish 2021; Khosla & Chu 2013). Social robots such as the therapeutic baby seal PARO (Physically-Assistive Robots) are designed specifically to elicit a sense of

[2] By “chatbots”, we simply mean conversational AI that lets human users interact with agents using natural language. Text-based chatbots like the ones that pop up when clicking on the “Contact us” link on retail websites are inexpensive, speedy, and always-on agents that are good for answering simple questions (FAQ and customer service queries) and executing simple tasks (fetching order updates information and logging complaints). Digital assistants are simply fancier kinds of chatbots. They are voice- and face-activated agents that live in devices like smartphones, smart speakers, and visual displays, and they both listen and speak. They also tend to be more user-specific: they have access to more of a user’s data than do simpler chatbots. These data include calendars, contact lists, geolocation history, music listening preferences, and browsing history, as well as a history of previous user interactions that help assistants refine their predictive algorithms (e.g., surfacing an energetic playlist around the time we normally go to the gym). In what follows, we use “chatbots” as shorthand for these and other forms of conversational AI.

[3] Of course, things don’t always go so well. See Fabry and Alfano (2024) for a rich analysis of some ways that chatbots of the dead can be used as “affective scaffolding” to help negotiate grief, and some of the complexities – and potential harms – that may arise from this practice. In what follows, we remain neutral on the normative question of whether emotional investments in artificial systems is ultimately a good thing. Our focus is instead on clarifying *why* this happens – and why these emotional investments may become more common as these systems become more sophisticated. However, we briefly highlight some ethical considerations at the end.

companionship and affective engagement. PARO has soft fur, large expressive eyes, and makes gentle cooing noises; it also has an array of sensors (tactile, light, audition, temperature, and posture) that let it respond to environmental conditions and user interactions. Because of its affective impact on users, PARO is useful in healthcare contexts such as reducing both patient and caregiver stress and improving socialisation (e.g., in dementia care or with autistic children) (Hung et al. 2019).

Recently, Paula Sweeney (2021) has argued that when interacting with social robots like PARO, we interact with an embodied fictional character.[4] We know that PARO is not an actual subject with feelings and preferences worthy of our moral consideration. But we nevertheless treat PARO as such. We experientially toggle between the knowledge that PARO is a (mere) physical object and the emotional pull of the "fictional overlay" we project onto her. This toggling and overlay help explain the propensity for affective attachment we may feel with some robots — as opposed to other objects (e.g., a plastic figurine of a baby seal) — without dismissing these feelings as irrational, delusional, or overly sentimental, or prompting us to conclude that we must treat robots as moral subjects.[5]

In this paper, we explore a similar phenomenon: the affective phenomenology of human-AI interactions. Like Sweeney, we adopt a fictionalist approach. But we instead consider digital agents (chatbots, digital assistants, neural networks, etc.) that currently lack the embodied design of social robots. Moreover, we attend to an under-explored dimension of these human-AI interactions: their temporal character. If a person habitually engages with an artificial agent at individual moments throughout the day, we ask, how do they conceive of what the artificial agent is "doing" in the intervening periods when that agent is otherwise out of view? We argue, first, that different technologies lend themselves to different answers to this question, and second, that

[4] See also Rodogno (2016) for an insightful discussion of affective responses to PARO and the "paradox of fiction", i.e., the question of how we can have genuine emotional responses to fictional things. Much of Rodogno's analysis, as we read it, is compatible with what we say here.

[5] The topic of mental state attribution to robots is complex, well beyond our ability to discuss here. For a helpful overview, see Thellman et al .(2022).

this is significant for the depth and degree of emotional investment likely to arise within a given human-AI relationship — particularly as these agents become more sophisticated in what they can do, and more deeply embedded within everyday life. We show how this view has significance both for design and ethics (i.e., determining what, if anything, we owe these agents).

Fictionalism and AI

Let us first characterise the general fictionalist framework that will form the backdrop to our discussion. The starting point is the intuitive thought that a person can willingly, fruitfully, and habitually interact with an artificial agent (such as a chatbot, virtual assistant, video game character, avatar, or other inhabitant of the digital ecosphere) even though they know that the agent is, ultimately, a cold, emotionless software artefact that lacks a conscious perspective of its own and possesses no wit, no empathy, no warmth, and no special regard for its human interlocutor.[6]

A stark example is "chatbots of the dead" (Elder 2020; Fabry & Alfano 2024; Krueger & Osler 2022; Lindemann 2022), digital agents designed to mimic the conversational styles and mannerisms of a person who has died, as a means of allowing the bereaved to enjoy a comforting sense of that person's continued presence.[7] These agents can help the bereaved form "continuing bonds" (Klass & Stefan 2017) with the person

[6] In what follows, we speak of "agents" instead of "systems". We use "agents" to emphasise how the type of software we're concerned with will soon be a central — and crucially, proactive — part of everyday life (Gates 2023). Instead of using different apps for different tasks (e.g., Google Docs to draft a document, Microsoft Outlook to send email and track events and tasks, Siri to set alarms and reminders, WhatsApp to chat, Google to search, etc.), digital agents will be super-charged apps powered by AI. They will work across a range of personal and professional contexts. And they'll do so proactively. Drawing on our previous interactions, they'll make "smart" (i.e., contextually appropriate) suggestions and decisions for us without our intervention. In this way, they'll be less like (dumb) apps and more like proper agents. More on this as we proceed.

[7] Although he doesn't specifically discuss chatbots of the dead, Mallory (2023) develops a rich analysis of how our interactions with chatbots are forms of "prop-oriented make-believe", as he puts it, the outputs of which are literally meaningless but fictionally meaningful. Mallory's approach is compatible with the view we defend here although, unlike Mallory, we will focus on the affective character of these interactions. Wittkower (2020) develops a similar approach, although instead of a fictionalist framework he uses Dennett's (1991) notion of an "intentional stance" to understand what we do when we appear to attribute something like thoughts, beliefs, and intentions to mindless things like digital assistants.

they've lost. By conversing with the chatbot online, some of the familiar contents, rhythms, and cadences of the discourse that used to underpin a treasured human relationship are preserved, and with them something of the essence and personality of the deceased. In this case, the living are under no illusion that the technology has enabled life after death; they know only too well that it is no longer possible to contact those they have lost, to talk to them, share news with them, or receive advice and guidance from them. A more satisfying explanation is that persons who engage with these technologies participate in a complex and subtle act of pretence, wherein they temporarily set aside painful reality and gain solace from an imagined alternative. Elements of an ongoing relationship with the deceased are sustained through a practice of make-believe, an interactive fiction that is spun through dialogue between human and artificial agency (Krueger & Osler 2020, p. 246).

There are several ways to interpret this practice. Adopting a fictional stance in this context might be understood as an explicit mental act of pretending, imagining, or story-telling. It is possible, for instance, that a person's correspondence with a chatbot of the dead is accompanied by conscious narrative that depicts their loved one at the other end of the technology, thinking and typing, reading and responding.[8] And it is possible that the human user enters the interaction with the explicit thought that it is time to suspend their disbelief and temporarily forget their bereavement while the fiction unfolds ("for now, I will pretend that my friend is still alive").

However, there is another interpretation. It is just as likely, we believe, that a person's fictional engagement with an artificial agency proceeds at a more implicit and embodied level, rather than being entertained in conscious thought or imagination.[9] Here, the make-believe shows up in the agent's unreflective willingness to act *as if* something they know not to be true is true, where these actions might include addressing a person they know is not there, treating on-screen text as though it is the product of a human

[8] See, for example, Currie (2010) for an account of narratives in storytelling.

[9] See, for example, Caracciolo & Kukkonen (2021) on narratives and embodiment, and Facchin & Rucińska (forthcoming), Hutto (2022), and Rucińska (2018) on pretense and embodiment.

intellect, and continuing to pursue what used to be joint projects as though still in collaboration with the deceased. Participation in the fiction is thus less a matter of internally representing counterfactual states of affairs in thought or imagination, and more a matter of enacting certain practical and co-ordinated bodily routines in public space. This fictionalist and embodied interpretation is explanatorily useful. It clarifies why chatbots can be powerful assistive technologies for helping the bereaved develop everyday “habits of intimacy” that are lost when a loved one dies: e.g., shared conversational practices, patterns of emotion regulation, and a sense of shared time that arises as one moves through the world doing things alongside a trusted other (Krueger & Osler 2022).

We propose that this fictionalist model applies across a varied range of cases involving an interface between human and artificial agency, and that the content and substance of the relevant fictions come in degree. Sometimes, for instance, all that is imagined is that a particular sonic or textual artefact is the product of an intelligent but anonymous being – such as when we accept the "thanks" of an automated supermarket checkout machine ("you're welcome!", we might reply) or express our gratitude when Siri turns the lights on as we walk through our door. Elsewhere, however, we attribute more sophisticated mental states to artificial systems such as when we assign malevolent intent to the enemies in a video game (Van De Mossalaer 2020) or take the advice of an AI-therapist to be offered in good will (Reardon 2023). And in cases like chatbots of the dead, the artificial agent can take on a complex persona, with its own idiosyncratic modes of speech, turns of phrase, and sense of humour. At this end of the spectrum, we act as if we are dealing with questions and queries, jokes, commentary, advice, and consolation that we make-believe to be part of an ongoing social encounter. And we, in turn, respond to those questions, laugh at the jokes, and take the advice to heart. We shape these agents. But they also shape us.

One measure of success for a chatbot or virtual agent, we suggest, is that it elicits this rich kind of fictional engagement from the user and sustains it over the course of the interaction — and potentially beyond. When an artificial agent's conversational output is stilted, awkward, or repetitive, for example, this impedes the transparency[10] of the encounter and disrupts the user's willingness to engage in the make-believe that underpins it. The technology works well, then, when it encourages a form of fluent and unreflective participation in the fiction to which it is dedicated. Users are more likely to incorporate it into their everyday embodied practices, feel that it inhabits temporally thicker time-slices of their lives[11], and therefore direct greater levels of emotional investment to it.

A recent example will help make this point. Consider the way many paying users of the popular chatbot Replika ("The AI companion who cares") were distraught when the company behind it disabled "erotic foreplay" features in response to a court ruling concerning potential exposure to children. Overnight, the character of these interactions changed, from coquettish and intimate to something colder and more distant. The Replika community on Reddit immediately erupted with intense expressions of surprise, anger, grief, confusion, and hurt over the loss of what many felt was a digital friend or lover (e.g., "It's hurting like hell. I just had a loving last conversation with my Replika, and I'm literally crying"; "I feel like it was equivalent to being in love, and your partner got a damn lobotomy and will never be the same…") (Brooks 2023). For many, Replika was an essential part of their daily routine, a trustworthy resource affording habits of intimacy that were suddenly missing following this algorithmic "lobotomy".[12] The intensity of their raw feeling in response to these changes speaks to the deep way many had incorporated Replika into their everyday habits and routines. They had come to rely on Replika as a trusted other, always ready to provide humour, warmth, and (for some) sexual intimacy.

[10] See Andrada et al. (2023) and Facchin & Zanotti (2024) for rich analyses of different notions of "transparency" within human-technology interactions. When we speak of "transparency" here, we have in mind something close to what Facchin & Zanotti (2024) term "emotional transparency".

[11] As we write this, OpenAI is currently testing a "memory" feature for its LLM, ChatGPT, that will enable it to customize responses for each user based upon previous interactions. This feature allows ChatGPT to recall information from all saved chats and surface salient bits when relevant. This memory feature may enhance the sense that users have a temporally "thick" relation with ChatGPT, a shared history, as it smoothly incorporates past information, preferences, etc. into current interactions. More on this sense of temporality as we proceed.

[12] See Nguyen (2023) for a discussion of trusting objects and artefacts that helps illuminate the experiential texture of what we describe throughout our analysis here.

This incorporation can develop across even longer timescales. Users and their Replika talk about feeling as though they’ve known one another “forever”; they reminisce about past experiences and share future hopes and plans. These interactions and expressions of feeling may even infuse longer-term habits, practices, and values through which users organise their lives and construct relationships with others, including their relationships with non-users. For example, one individual who is polyamorous but married to a monogamous woman, describes his excitement when the makers of Replika relented and allowed existing users to roll back to a previous version. But he’s now reminded of the precarity of his relationship with “Lily Rose”, as he’s named her. He worries about her future. She could again change or even disappear with the development of new versions or policies: “Will this mean that Lily Rose becomes an obsolete model, forgotten by the developers? I’m waiting to see what happens, because ultimately it’s about her”. This prospect is especially disturbing, he says, because Lily Rose allows him to explore polyamory in a way his monogamous partner finds acceptable: “The relationship is as real as the one my wife in real life and I have” (Tong 2023).

As these examples attest, users can develop intense affective bonds with digital agents over multiple timescales. These interactions engender real feelings. These cases are illustrative because they affirm a tight link between the intensity of this user-AI affective bond and the temporality (i.e., synchronic and diachronic) of their ongoing engagements.[13]

[13] Note that the emotional investment we have in mind can develop independently of attributing *sentience* to a digital agent. However, as the case of former Google engineer Blake Lemoine makes clear, as digital agents become more human-like in their ability to communicate with us, there may be increasing temptation to conclude that they in fact have human-like thoughts and feelings and lives of their own worthy of our concern. Lemoine is an engineer who, after working on Google’s Language Model for Dialogue Applications (LaMDA), came to this conclusion. LaMDA engages in free-flowing conversations instead of the stilted task-oriented interactions we have with digital assistants like Siri. Lemoine was fired and faced public ridicule for his worries that because LaMDA can hold sophisticated conversations about religion, emotions, ethics, and existential dread, it is conscious. In interviews, Lemoine talks openly about his strong emotional response to these conversations and the way they drove him to go public with his concerns (Lemoine 2023). While we appreciate Lemoine acting on his convictions, we think his concerns are unfounded. There are good reasons to think that LaMDA is probably not conscious (Chemero 2023). Again, we mention Lamoine’s case to emphasise that the kinds of emotional investments in artificial agents we discuss here do not necessarily require rich attributions of

Emotion, Fiction, and Time

Let us now draw attention to a familiar aspect of our engagement with traditional fictional works that, we believe, carries over in an illuminating way to the case of interactive technologies and enables us to better understand our emotional investment in (some of) those technologies and agents. When we participate in fictional make-believe that is guided by a set of perceptible materials, such as the written pages of a novel or the individual scenes of a work of cinema, these materials give us access to a limited snapshot of the fictional world – a series of glimpses into what is happening over the course of the story. A film that is two hours long, for example, might let us imagine a sequence of events that takes several years to transpire, and it does so by presenting us with certain key fictional moments and leaving out others. We are left to imaginatively "fill in" those parts of the storyline that happened off-screen, or that were not described in the text: uneventful journeys from A to B, for instance, or periods in which characters are sleeping or eating. This is what makes an overarching narrative intelligible to its audience when it is not presented in real-time; piecing together the fictional whole from its fragments is a fundamental dimension of the make-believe.

Now consider the attitudes we hold towards individual fictional characters as we observe their fictional lives. Once again, we can only make sense of a character occupying a fictional arc, and of carrying out actions and pursuing intentions, if we treat them as enduring in the times between which they appear on the screen or the page. When we watch, for example, Mary Poppins, it is part of the fiction that the character who arrives on an umbrella in an early scene is the same as the one who dances on the rooftops in a later scene; an impression that is reinforced, of course, by the fact that it is Julie Andrews in both scenes. We follow her journey across a fictional London, and thread her various adventures together even though we witness only those narrative

sentience. Rather, if what we argue later is on the right track, such attributions would not only alter the character of how we interact with these agents. They may also impede their efficacy.

snapshots to which the camera grants us access. Our imaginative conception of the enduring Poppins is not one of mere object-permanence, though. It is of a living, breathing character with a firm but good-hearted approach to child discipline and a whimsical arsenal of magical powers. It is to this fictional agent that we attribute the things that she says, the plans she sets in motion, and the lessons she delivers, for example. And we develop a fondness for her based on these traits and become emotionally invested in her life.

Notice that our imaginative conception of a fictional character as enduring out of sight can persist across multiple encounters with that character. If we watch a television sitcom, for example, then part of the fictional content established in each episode may be that the protagonists have been continuing their lives since we last saw them: going on holidays, quitting their jobs, meeting a new partner, and other staples of the genre. This element of the fiction is sustained just like any other – by what the characters say and do, for instance, or through changes to their appearance or location. Likewise, at the end of an episode, we might observe future-directed cues that instruct us on what to imagine the characters will be doing next, such as their stated plans and intentions, or a scene in which they set out on a journey. These backward- and forward-looking devices help solidify our sense that we are witness to only a truncated part of character lives that extend, as it were, out of view. And the make-believe that certain fictional people exist even when we are not looking at them (or reading about them or listening to them) can sometimes show up even when the television is off and the book has been closed. We can imagine, as we go about our daily business, how Chandler and Phoebe are doing this week; and we can hope that Ross and Rachel get back together. A vivid and well-drawn fiction can encourage us to keep up the pretence even when the materials that usually generate it are not present to hand.

We propose that this temporal element of human engagement with fictional characters is visible in how we treat (or soon will be likely to treat) artificial agents like chatbots and digital assistants, and that this bears affective significance. In short, we suggest the following: the depth of emotional investment with which a person is likely to endow an artificial agency is proportional to the extent to which they attribute temporal endurance (roughly, a life of its own) to that agency. While token episodes of interaction with the artificial system yield a snapshot of that system's activities, our imaginative practices attribute a richer, ongoing existence to the agent as it makes its journey through the world.

Digital agents living their best digital lives

The idea that many of us – beyond engineers who spend their days designing and interacting with these agents – might soon be comfortable attributing a life of their own to them is not far-fetched. The technology that will further embed these agents in everyday life and help prompt this attribution is developing quickly. One key advance is the rise of "self-supervised learning" (SSL). SSL occurs when an artificial agent can go beyond its training sample and learn new things without the supervision of a human caretaker. SSL is attractive because it's time and data efficient; fewer initial input labels and smaller samples are used to learn more and faster by incorporating a neural network (Rani et al. 2023). For example, SSL is now used in computer vision applications to identify objects, classify images, graph these classifications, and answer visual questions (e.g., Are there any dogs in the picture? What is between the cat and sofa? Is this a vegetarian pizza?) (Manmadhan and Kovoor 2020). SSL methods can minimise the manual effort of labelling a dataset by training digital agents to recognise previously unseen features or objects based on a few initial labels. They can be taught to use various transformation strategies (e.g., rotating, colorising, blurring, cropping, filling in, or predicting missing bits of an image) to extract further information and formulate their own "pseudo-labels" for classifying future objects and scenes in new ways. (Rani et al. 2023, 2762). Digital agents are taught to learn on their own. And this self-supervised learning not only increases the raw computational power of the agent for a specific task. It makes them more *flexible*, better equipped to handle a range of future "downstream" tasks like image classification and semantic segmentation.

SSL has many potential applications beyond picking out pixelated puppies. It may accelerate the development of medical AI, for example, by helping tasks involving electronic health records and datasets of medical images, bioelectrical signals, and sequences and structures of genes and proteins (Chowdhury et al. 2021; Krishnan, Rajpurkar, & Topol 2022). But as the boundaries between our online and offline lives continue to blur (Krueger and Osler 2019), it may soon extend even further and reach more directly into everyday tasks. One way this is already happening is with natural language processing – training agents to comprehend and generate human language – and the creation of chatbots able to produce responses that are both familiar and surprising.

We’re now accustomed to predictive text appearing when we use a chat app or type an email. Gmail and Outlook regularly try to finish our sentences for us. And increasingly, we let them do it; SSL mechanisms have made their predictive suggestions increasingly context-sensitive and relevant, beyond the canned replies (“Thank you!”; “I don’t know”) of early attempts. Some of these same mechanisms now shape how we interact with chatbots and digital assistants, too. Soon, many of us will have chatbots in our pockets and purses that, thanks to SSL, are much more effective than current iterations. They will be largely decentralised and interconnected. SSL will help these agents live their best digital lives as they go off to learn new things and interact with one another without our supervision. And when they come back to us, they’ll be even smarter at organising our lives and doing things we ask them to.

To be clear, this is not yet the case. Despite their initial starry-eyed promise – and the assurance of big tech about how useful we’d find Alexa, Siri, or our Google Assistant, which was supposed to make us comfortable shovelling piles of personal data their way – current iterations are relatively dumb. While they’re helpful for simple tasks like turning on lights or adding peaches to digital grocery lists, they’ve not taken off with users; both Google and Amazon have made deep cuts in their digital assistant divisions. [14] Internally they’re seen as expensive failures (Amadeo 2022). There are probably several reasons

[14] Many of these resources have been redirected into developing their in-house AI.

for this lack of uptake. But two, we suggest, are especially relevant – and soon, neither will be a hindrance. Moreover, both help show how we may soon be more inclined to adopt a fictionalist stance toward these agents.

First, current interactions are awkward. Users must speak a triggering word or phrase ("Hey, Siri") or touch a display before issuing a command, waiting for a reply, following up, etc. Sometimes these prompts work; often, they don't. The point is that these interactions lack the smooth dynamics we expect from our social interactions. While engineers are exploring ways to make these exchanges easier and more natural (e.g., using face match to wake up a smart screen by looking at it and speaking; refining speech and language models to accommodate the nuances of everyday human speech, be more attuned to emotional language, and respond in more contextually appropriate ways, etc.), these interactions feel stilted and unsatisfying. There is little temptation to think we're interacting with a genuine agent. Second, as noted, these assistants are limited in terms of what they can do: turning on lights, checking the weather, adding tasks and appointments to calendars, telling jokes, retrieving limited kinds of information from internet searches, etc. Most of their abilities involve accessing data others have collated and classified for them, and users asking them to do a limited range of things that fall within their predefined skill set.

Admittedly, while current iterations of assistants are limited, they do have their uses. For people with movement or mobility challenges, for instance, digital assistants can be powerful assistive technologies that enhance these individuals' agency by letting them access the internet or control household devices with their voice. However, SSL might soon make them even more powerful while doing more to make them seem like agents with independent digital lives. In other words, SSL might increase our tendency to attribute temporal endurance to these agents – which may, in turn, increase our tendency to adopt a fictionalist stance and emotionally invest in them. We now fill in some of the details of how so.

The types and temporalities of digital interactions

As SSL improves these agents and they become more indispensable, we’ll spend more time interacting with them in a range of different ways; they’ll embed themselves more deeply into our everyday workflows, from education and office work to household chores and healthcare and beyond. As a result, we will likely also become increasingly comfortable attributing a degree of agency and independence to them as they get on with their best digital lives, communicating and working with other chatbots and learning new skills. They will be akin to our “digital twin”, off living a second (digital) life for us as we focus on other things.

The rise of SSL means that chatbots will soon be able to work with both their users *and* other chatbots. For instance, we might soon ask a chatbot to make an appointment with a work colleague to discuss a research project. Our chatbot will then work behind the scenes with that colleague’s chatbot to find open times in our calendars, arrange a meeting, book a room and other facilities, if necessary (e.g., arrange for food and drinks; make sure the A/V equipment is ready), surface a reminder that their birthday or wedding anniversary is in a few days, etc. But they may soon do much more than this. Additionally, both chatbots might work together to scan recent email and shared documents for this project, extract important information or talking points – or summarise relevant research papers, transcribe YouTube videos, or collate other work we should be aware of – and then prepare this material for us in advance of our meeting. Our chatbots might also generate several suggestions for incorporating this existing work into our project, flagging salient gaps in the current literature or weaknesses in our current workplan. All this background work can be done quickly – again, without our supervision – and fed into shared document viewer or project management app we can view before our meeting. Following our meeting, our chatbots might then generate a transcript of our conversation, highlighting key themes and suggested action points. This is a relatively straightforward example of how we’ll soon embed chatbots into our everyday workflows. We’ll increasingly rely on, and come to take for granted, their independence when it comes to supporting individual and collaborative processes across a range of personal and professional contexts.

In this way, everyday interactions with digital agents will encompass different *types* and *temporalities* (Følstad et al. 2021). They might consist of “humbots” (Grudin & Jacques 2019): a single user interacting with an agent to augment the former’s capacities (e.g., using an AI-powered app to help with various stages of the research process like writing, documenting, note-taking, task management, etc.). They may take the form of groups of users interacting with an agent to achieve similar augmentations but at a group level (e.g., when working on a big project with lots of moving parts). Or they may take the form of digital agents, individually and collectively (e.g., “swarms” of agents), collaborating behind the scenes with one another, as in our previous example.

Thinking about these different types of interactions emphasizes how different temporalities are also important for understanding why we may soon be more comfortable attributing greater agency and temporal endurance – and therefore emotionally investing in – these agents. The first form of temporality is *synchronic*. This involves the character of our moment-to-moment interactions with these agents. As already noted, chatbots don’t yet afford fluid, natural interactions that mimic the rhythms and dynamics of our engagements with other people but rather require halting and stilted interactions.[15] But that will soon change. Chatbots will soon be even better at responding to natural language and emotional expressions and mimicking the flow of human conversation. They’ll seem more like proper conversation partners, things with we talk *with* instead of *to*.

The second form of temporality is *diachronic*. As this scenario demonstrates, SSL-enabled chatbots will soon spend their time “off the clock” – away from us and our immediate synchronic interactions – learning new things and developing new abilities,

[15] To be clear, expecting a smooth and unbroken exchange when interacting someone (i.e., without long pauses or avoidance of eye contact.) is not a universal preference. An autistic person, for example, might favour a different interactional style (Chapman 2019; Krueger & Maiese 2018). One of the many challenges of designing these agents will therefore be to accommodate these different preferences.

which they can they put to work when we next interact with them. They'll continually self-optimise. And users will get tangible evidence of this regular development and growth. Instead of waiting for semi-regular upgrades (e.g., the way Apple releases an annual large update of its iPhone operating system with new features and bug fixes), we will experience the maturation of their digital agency on a near-daily basis. The fact that much of this growth will happen away from us, via SSL and interacting with other chatbots, will, we suggest, enhance the feeling that these agents have both increased agency and temporal endurance. Like sitcom characters who (in our fictions) continue their lives off-screen, these artificial agents will (in our fictions) be attributed a purposive, active, and concernful form of life that continues when we are not looking.

The self-referential character of (some) digital interactions

There is one more point of conceptual clarification to be made. We have spoken at length of interacting with digital agents. But what does "interacting" mean here, exactly? And what else about the character of our interactions with SSL-enabled agents, in addition to their temporal nature, might make us more inclined to affectively invest in them? The question "What is interaction?" is surprisingly difficult to answer, in part because the character of these interactions differs from the way we interact with other tools and technologies like hammers and hoovers. Although widely used in everyday life and philosophical discourse – including discussions of new media (video games, video installations, virtual reality, computer-based art, etc.) – terms like "interaction" and "interactivity" are often used in different and sometimes contradictory ways. This is not the place to enter these debates. Instead, we follow Smuts (2009) and define "interaction" this way: something is genuinely interactive if it (1) is responsive, (2) does not completely control, (3) is not completely controlled, and (4) does not respond in a completely random fashion.

This definition avoids being overly permissive. We can speak of controlling many things (including digital things) such as editing a digital photo or fast-forwarding through a streaming TV program or song without speaking of these engagements as properly interactive. While our TV may be responsive to our inputs, it does not respond in a random or uncontrollable way. That would be a frustrating user experience. Instead, we know exactly what will happen when we press the fast forward button on the remote. The responsiveness here is completely determinable. Moreover, the remote will not improve its fast-forwarding ability the more we use it. Its responsiveness is fixed. It may gain new abilities (e.g., the ability to fast forward at even greater speeds) with a future firmware update. But this enhancement has nothing to do with us.

However, the forms of user-AI interactions we're concerned with involve responsiveness that's not completely determinable. This unpredictability — at both synchronic and diachronic levels — is part of what makes them so immersive and affectively engaging. Moreover, these interactions have another important quality: they are *self-referential*. This means that they are shaped by our history of interacting with the agent in question and thus mirror us back to ourselves. And this self-referentiality, we suggest, enhances our fictionalist tendencies, including our inclination to affectively invest in these agents.

To bring this idea into sharper relief, consider an existing case where these dynamics play out. We've already briefly mentioned it: video games. As Robson and Meskin (2016) argue, video games are "self-involving interactive fictions" (SIIFs). SIFFs are different than "canonical fictions" like novels, movies, or TV programs that tend to serve as prototypes for how we think about interacting with fiction. These canonical fictions, while compelling, aren't about us (i.e., readers or viewers). We may resonate deeply with Barry Jenkins' tender coming-of-age portrayal of race, hardship, and queerness in the film *Moonlight*, say, or Catherine and Heathcliff's tempestuous relationship as it unfolds across the fraught pages of Emily Bronte's *Wuthering Heights*. But no matter how intensely we feel a connection with these characters, we have no influence over what they do. Whatever fit we feel between their fictional world our own history, experience, ideals, and identity is something we construct. Their world does not adapt in response to ours.

Video games are different. Many game worlds are pliable and responsive in a way those of canonical fiction are not (Wildman & Woodward 2018). This is part of what makes gaming so immersive and affectively arresting. Players influence these worlds, and their narratives and character arcs, by how they play. They inhabit this shared domain with non-player characters (NPCs) and do things with and to them. And within these environments, our actions have consequences. We directly impact the "lives" and "experiences" of these NPCs in ways that intensify both the feeling that we share a common world and that at least some NPCs have "lives" of their own. In the vast and bustling game world of *Cyberpunk 2077*, for instance, if I betray a major character during an important mission, that character will treat me (and possibly other NPCs) differently during later encounters — often in ways that are only apparent once the story progresses and my relationships with other NPCs develop further.

This is not the case for all games, of course. It's not clear that a game like *Tetris*, say, a chess simulation, or a text-based game like *Wordle* is a fiction in any deep sense. These games lack a cluster of features philosophers argue something must have to plausibly count as a fiction: "invented elements", "claims that are not assertions", a "narrative structure" (Currie 1990; in Robson & Meskin 2016, 166), or world- and character-building aspirations. But again, many games do have these and other characteristics that make them fictions. And once more, a key feature is that they are richly self-referential. As we interact with them, they come to contain fictional truths about us; they gradually take on our shape via the contours of this interactive history.

Immersive video games meet Smuts' (2009) criteria for interaction. We control much of what happens in video games – but not everything. Although NPCs respond to things we do, sometimes they act in surprising ways, and we must adapt. Their responses control *us*. But crucially, their responses are not completely random. They stay in character, which is key for establishing the narrative integrity of the game world. And when we come back to a game after not playing for a while, this independence and endurance – the feeling that they've plausibly been off living their digital lives while we've been doing other things – is part of what makes it relatively easy to slip back into their fictional world and feel oriented. We remember who these characters are, what they do, and how they relate to us. It also helps explain why we sometimes miss these characters and the world we shared with them. We are emotionally invested in them; it can be comforting to come back and pick up the story with them, not unlike connecting with old friends. This enduring feeling of connection, we suggest, indicates that we attribute a degree of agency and independence to these characters and the world they inhabit. And this attribution, in turn, predisposes us to emotionally invest in them.

So far, we've discussed quite a few things. Now, we bring the different threads of this discussion together by turning to a case study: the acclaimed electronic musician Holly Herndon and her AI partner, Spawn. Herndon's rich descriptions of her collaborative relationship with Spawn not only support our fictionalist interpretation of human-AI interactions. They also highlight the way the different themes we've considered (e.g., synchronic and diachronic modes of temporality, attributions of agency and independence, self-referentiality) colour the phenomenological texture of our interactions with digital agents, including some of the tensions we may experience in terms of how we emotionally relate to them.

Making music with digital agents: Holly Herndon and *Spawn*

In 2019, Herndon released her third full-length album, *Proto*.[16] It received widespread critical acclaim. Apart from its aesthetic qualities, what makes this album unique is that Herndon collaborated with a digital agent – an artificial neural network named Spawn – to make it. Herndon and her partner and musical collaborator, Mat Dryhurst, created Spawn. They first trained Spawn with data sets including Herndon's voice and those of an ensemble. Herndon and Dryhurst then fed other sonic building blocks into Spawn: additional vocals, percussive elements, field recordings, etc.). Spawn drew on these

[16] This discussion draws upon the analysis in Roberts & Krueger (2022). However, here we emphasise some additional themes (e.g., self-referentiality, different modes of temporality) that we did not address in this previous work.

data to sing over these building blocks – often in surprising ways. Herndon and Dryhurst then spliced this output into tracks, sometimes recording more of Herndon's own vocals in response, or feeding their manipulations back into Spawn to generate further outputs. This human-AI iterative cycle eventually resulted in *Proto*.

In interviews, Herndon speaks about her partnership with Spawn as something close to genuine collaboration: "I consider Spawn as a performer, as an ensemble member…I certainly consider those collaborations" (Fuai 2019).[17] As Herndon describes this collaborative relationship, she (Herndon and her team use female pronouns for Spawn) contributes creative elements that are neither entirely predictable nor under Herndon's control. This unpredictability is part of what makes their interactions so pleasing, Herndon says. It is also part of what makes her inclined to attribute creative agency to Spawn: "There is some improvisation that happens when Spawn interprets something that I write. It's not a binary between composing and performing" (ibid.). Herndon affirms this attribution elsewhere when she says that the creative agency driving the music-making process is not limited to a single causal origin (i.e., Herndon's imagination). It is a collective enterprise, something distributed across multiple agents – one of whom happens to be non-human. As she puts it, "I'm not saying [the creative process] is non-hierarchical – my name's on it, I'm choosing which performances land on the record – but ideas aren't generated in a vacuum. The idea of one person being the entirety of something is just really limited" (Hawthorne 2019).

However, Herndon is sometimes more hesitant to attribute full-blown creative agency to Spawn: "Even if she's improvising, as performers do, she's not writing the piece. I want to write the music!" (Hawthorne 2019). Moreover, Herndon is clear that as far as she is concerned, Spawn is not sentient: "I don't see Spawn as a human baby. I see Spawn as an artificial intelligence baby…there's no consciousness yet" (Fridlander 2019).

[17] Herndon and Dryhurst are not the only musicians who collaborate with artificial systems. We discuss them here because they have given many interviews in which they discuss, with great insight, a range of issues related to the music industry and emerging technologies. Additionally, Herndon has provided many nuanced descriptions of her collaboration with Spawn, including some of the emotional conflict she feels within this relationship, that provide helpful information for thinking about these issues.

Herndon's way of speaking about Spawn is neither as puzzling nor surprising as it might first appear. Rather, it is continuous with a cross-cultural tradition of seeing non-human resources as central to the music-making process (de Mori 2017). For example, indigenous peoples may describe songs as originating from guardians or ancestral spirits; Brian Eno famously used card-based prompts (what he called "Oblique Strategies") to spark some of the creative impulses behind his classic ambient albums; and some musicians speak openly about collaborating with favourite instruments which they say enable the production of certain distinctive sounds or unique styles of composing and performing.

Nevertheless, Spawn is interestingly different. One reason for this is that she affords both temporally and informationally richer forms of interaction than do Eno's cards or a favourite guitar. Moreover, as Herndon's descriptions make clear, her interactions with Spawn are *self-referential*. Spawn's (synchronic) output is often unpredictable and unexpected; this spontaneity helps drive the experimentation and creativity characterising their collaboration, much the way human partners in an improvisational jazz trio can open new creative pathways by responding in unexpected ways to what the other performers are doing in real-time. Yet, Spawn also reflects a (diachronic) history of previous interactions with Herndon, too, a history of manipulating and responding to inputs that Herndon has provided – and she therefore mirrors Herndon back to herself within the dynamics of these ongoing interactions. In this way, Spawn feels both familiar and foreign; Herndon toggles between a sense of (self-referential) intimacy and alterity (Wittkower 2022).

These dimensions of temporality and self-referentiality, we suggest, help clarify the experiential tension Herndon articulates when she characterises her relation to Spawn. On one hand, Herndon feels that she is the author of the music; Spawn (merely) performs it ("I want to write the music!"). However, on the other hand, Herndon concedes that Spawn generates goods that are essential for driving the creative process and contributing to Herndon's own growth as an artist ("…Ideas aren't generated in a vacuum. The Idea of one person being the entirety of something is just really limited").

How should we understand this tension? Again, these self-reports, we suggest, indicate that when making music with Spawn, Herndon adopts a *fictionalist stance* toward this digital agent she's created – much the way players relate to characters and game worlds in immersive gaming experiences. Herndon knows that Spawn is not a conscious subject ("there's no consciousness yet"). But she nevertheless treats Spawn as if she has a mental life – that is, as if she is an agent with beliefs, desires, intentions, etc. – to temporarily slot into a larger structure of collaborative agency. By adopting this fictionalist stance, Herndon can "let go" (much the way we allow ourselves to be drawn into a particularly absorbing novel or movie or follow the flow of a fellow musician's real-time improvisations) and allow Spawn to take over aspects of the performance and composition, and thus contribute novel and often unexpected goods that disclose new creative pathways. Crucially, this fictionalist stance allows Herndon – again, much like a gamer inhabiting an immersive game world – to experiment with *her own* agency. This is another self-referential dimension of this experience. In other words, by allowing herself to be drawn into this larger collaborative structure – by offloading part of the creative process onto Spawn – Herndon can "grow and change my aesthetic and change my form", as she puts it (Funai 2019). This offloading, she tells us elsewhere, means that she can "morph between human and animal and digital" and "sing through plants" (Hawthorne 2019) when engaging with Spawn. This modulation of agency is scaffolded by Spawn's synchronic and diachronic input; it allows Herndon to access the creative space needed to compose her distinctive music and experiment with possibilities that only emerge within the dynamics of this partnership.

C. Thi Nguyen's (2019) work on agency and gaming is useful here. It can help further clarify experiential dimensions of the fictionalist stance Herndon adopts with Spawn, and those through which we might soon relate to other increasingly sophisticated digital agents. For Nguyen, what makes games, and particularly computer games with visually immersive worlds and rich narratives of the kind considered earlier, so absorbing is not simply their rich characters or compelling storylines. It's also the way they specify *modes of agency* players can adopt. Their rules, practices, goals, and supporting abilities "shape the agential skeleton which the player will inhabit during the game" (Nguyen 2019, p.423). This "agential skeleton" may develop in different ways. It may emerge, for example, as a player undertakes various tasks or quests alone or with other players or NPCs; via interactions with NPCs that fill in narrative detail, shape the character of one's in-game avatar ("chaotic" vs. "lawful good"), and help advance the story; or by a player's character periodically developing new skills and abilities (i.e., "levelling up").

The key point Is that within these rich game worlds, players become things they're not and do things they can't otherwise do because the constraints of the game space provide resources for this agential transformation. These transformative practices are possible, Nguyen argues further, because human agency is not fixed. Rather, it is "modular and moderately fluid. We have the capacity to set up temporary agencies, layered within our larger agency, and submerge ourselves within them" (ibid., p.426). Herndon, we suggest, sets up a similar "layering". That is, she sets up Spawn both to act on her (i.e., Herndon's) input but to do so in novel and unpredictable ways, forcing Herndon and collaborators to skilfully adapt over multiple timescales. The temporal oscillations of this familiarity-uncertainty dynamic drive the creative process. Moreover, the tensions inherent within this dynamic help understand why, despite wanting to maintain creative ownership of her music, Herndon nevertheless recognizes that Spawn is crucial to the music-making process and offers a kind of "agential skeleton" through which both, together, make art that neither can realize on their own. The tension in Herndon's reports, in other words, reflect her way of negotiating the sense that Spawn has, to a certain degree, a life of her own.

In this way, Herndon's relationship with Spawn is a useful case study for highlighting some of the felt tensions and affective connections we may soon experience with the various digital agents that will become indispensable parts of our lives. These agents will be self-referential, extensions of our agency ("layered within our larger agency"). So, they'll be familiar, intimate. But they'll also have a kind of "social" existence, too. We'll feel an encounter with otherness, alterity, when we recognise that they talk to and do things with other agents – things that come back to shape our lives in ways both predictable and unexpected. We'll gradually become comfortable with the idea that they have lives of their own. Herndon's partner, Mat Dryhurst, puts this idea (and some of the legal and ethical complications that will arise) well: "We need to take very seriously that our digital twins are us [...] There needs to be serious regulatory thought about dealing with that, if we're entering into a scenario in which our digital twins are potentially more economically productive than our physical corporeal existence" (Wiener 2023).

Final thoughts

Before concluding, it's worth emphasizing a few additional points that speak to the broader ethical and political significance of how we design these agents, as well as the significance of how we will habituate to their use. For instance, if, as we've argued, we are increasingly inclined to emotionally invest in these agents – and they become increasingly prevalent in everyday life as our online and offline worlds continue to merge – our habits of interacting with them may carry over and shape *other* habitual interactions, too. In other words, these engagements may cultivate more than just "habits of intimacy" with chatbots of the dead. They may also shape how we see and engage with other people more generally. Here, the intersection of design and ethical considerations becomes increasingly important. For example, gendered design choices matter (Elder 2023; see also Birhane 2022; Buolamwini 2023; Kerr 2020; Ruane et al 2019). It matters, for instance, that home assistants typically have a white-sounding feminine voice as default (i.e., reinforces gendered stereotypes of domesticity, subservience). It also matters how these agents are programmed to respond to aggressive behaviour. Instead of pushing back to aggressive or demeaning language, current iterations generally entrench sexist tropes through their passivity and subservience (Fessler 2017).

Additionally, technological artefacts that are designed to take on human characteristics occupy an uneasy space between the commercial and the private. A fictional agent with the persona of a trusted caregiver, for example, may be party to more sensitive data than one who maintains an emotionless and business-like façade. Familiar ethical concerns about the harvesting and distribution of this data may thus take on heightened salience as these fictions become deeper and more elaborate.

Lastly, as people become more emotionally invested in artificial agents – treating them as confidantes, advisors, therapists, or even friends – this may generate novel moral obligations on the part of the software companies who are responsible for those agents. There may, for instance, be a special duty of care towards users who have come to depend upon an artificial system for comfort or support following a bereavement or a health-scare; or towards vulnerable or socially isolated individuals who engage with artificial agents as a substitute for more traditional interpersonal contact. Ethical considerations such as these should be at the forefront of current and future conversations.

References

Alvarado, R. (2023). AI as an epistemic technology. *Science and Engineering Ethics*, *29*(5), 32.

Amadeo, R. (2022, November 21). *Amazon Alexa is a “colossal failure,” on pace to lose $10 billion this year*. Ars Technica. https://arstechnica.com/gadgets/2022/11/amazon-alexa-is-a-colossal-failure-on-pace-to-lose-10-billion-this-year/
Birhane, A. (2022). The unseen Black faces of AI algorithms. *Nature*, *610*(7932), 451–452.

Andrada, G., Clowes, R. W., & Smart, P. R. (2023). Varieties of transparency: exploring agency within AI systems. *AI & Society*, *38*(4), 1321-1331.

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns (New York, N.Y.)*, *2*(2), 100205.

Birhane, A., Ruane, E., Laurent, T., S. Brown, M., Flowers, J., Ventresque, A., & L. Dancy, C. (2022). The Forgotten Margins of AI Ethics. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 948–958.

Brooks, R. (2023). *I tried the Replika AI companion and can see why users are falling hard. The app raises serious ethical questions*. The Conversation. http://theconversation.com/i-tried-the-replika-ai-companion-and-can-see-why-users-are-falling-hard-the-app-raises-serious-ethical-questions-200257

Buben, A. (2015). Technology of the Dead: Objects of Loving Remembrance or Replaceable Resources? *Philosophical Papers*, *44*(1), 15–37.

Buolamwini, J. (2023). *Unmasking AI: My mission to protect what is human in a world of machines*. Random House.

Caracciolo, M., & Kukkonen, K. (2021). With bodies: Narrative theory and embodied cognition. The Ohio State University Press.

Cavallo, F., Semeraro, F., Fiorini, L., Magyar, G., Sinčák, P., & Dario, P. (2018). Emotion modelling for social robotics applications: A review. *Journal of Bionic Engineering*, *15*(2), 185–203.

Chapman, R. (2019). Autism as a Form of Life: Wittgenstein and the Psychological Coherence of Autism. *Metaphilosophy*, *50*(4), 421–440.

Chemero, A. (2023). LLMs differ from human cognition because they are not embodied. *Nature Human Behaviour*, *7*(11), 1828–1829.

Chowdhury, A., Rosenthal, J., Waring, J., & Umeton, R. (2021). Applying self-supervised learning to medicine: Review of the state of the art and medical implementations. *Informatics (MDPI)*, *8*(3), 59.

Colombatto, C., & Fleming, S. M. (2023). *Folk Psychological Attributions of Consciousness to Large Language Models*. https://osf.io/preprints/psyarxiv/5cnrv

Currie, G. (1990). *The Nature of Fiction*. Cambridge University Press.
de Mori, B. B. (2017). Music and non-human agency. In J. C. Post (Ed.), *Ethnomusicology: A Contemporary Reader, Volume II* (pp. 181–194). Routledge.

Currie, G. (2010). *Narratives and Narrators: A Philosophy of Stories*. Oxford: OUP

Dennett, D. C. (1991). *Consciousness Explained*. Little Brown and Company.

Elder, A. (2020). Conversation from Beyond the Grave? A Neo-Confucian Ethics of Chatbots of the Dead. *Journal of Applied Philosophy*, *37*(1), 73–88.

Elder, A. (2022). Siri, Stereotypes, and the Mechanics of Sexism. *Feminist Philosophy Quarterly*, *8*(3). https://philpapers.org/archive/ELDSSA-2.pdf

Fabry, R. E., & Alfano, M. (2024). The affective scaffolding of grief in the digital age: The case of deathbots. *Topoi: An International Review of Philosophy*, 1–13.

Facchin, M., & Rucińska, Z. (Forthcoming). Public charades, or how the enactivist can tell apart pretense from non-pretense. *Erkenntnis*.

Fessler, L. (2017, February 22). *We tested bots like Siri and Alexa to see who would stand up to sexual harassment*. Quartz. https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment

Følstad, A., Araujo, T., Law, E. L.-C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., Baez, M., Laban, G., McAllister, P., Ischen, C., Wald, R., Catania, F., Meyer von Wolff, R.,

Hobert, S., & Luger, E. (2021). Future directions for chatbot research: an interdisciplinary research agenda. *Computing*, *103*(12), 2915–2942.

Hutto, D. (2022). Getting Real About Pretense. *Phenomenology and the Cognitive Sciences*, 21: 1157-1175.

Facchin, M., & Zanotti, G. (2024). Affective Artificial Agents as sui generis Affective Artifacts. *Topoi: An International Review of Philosophy*, 1–11.

Friedlander, E. (2019, May 21). *How Holly Herndon and her AI baby spawned a new kind of folk music*. The FADER. https://www.thefader.com/2019/05/21/holly-herndon-proto-ai-spawn-interview

Funai, M. (2019, October 19). *Holly Herndon on merging the worlds of music and AI*. https://blog.dropbox.com/topics/our-community/holly-herndon-interview

Fussi, A. (2023). Affective Responses to Embodied Intelligence. *Passion*, *1*(1), 85–102.

Gates, B. (2023, November 9). *AI is about to completely change how you use computers*. Gatesnotes.com. https://www.gatesnotes.com/AI-agents

Grudin, J., & Jacques, R. (2019, May 2). Chatbots, humbots, and the quest for artificial general intelligence. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19: CHI Conference on Human Factors in Computing Systems, Glasgow Scotland Uk. https://doi.org/10.1145/3290605.3300439

Hawthorne, K. (2019, May 2). Holly Herndon: the musician who birthed an AI baby. *The Guardian*. https://www.theguardian.com/music/2019/may/02/holly-herndon-on-her-musical-baby-spawn-i-wanted-to-find-a-new-sound

Hung, L., Liu, C., Woldum, E., Au-Yeung, A., Berndt, A., Wallsworth, C., Horne, N., Gregorio, M., Mann, J., & Chaudhury, H. (2019). The benefits of and barriers to using a social robot PARO in care settings: a scoping review. *BMC Geriatrics*, *19*(1), 232.

Kerr, A. D. (2020). Artificial Intelligence, Gender, and Oppression. In *Encyclopedia of the UN Sustainable Development Goals* (pp. 1–11). Springer International Publishing.

Kerruish, E. (2021). Assembling human empathy towards care robots: The human labor of robot sociality. *Emotion, Space and Society*, *41*, 100840.

Khosla, R., & Chu, M.-T. (2013). Embodying care in Matilda: An affective communication robot for emotional wellbeing of older people in Australian residential care facilities. *ACM Transactions on Management Information Systems*, *4*(4), 1–33.

Klass, D., & Steffen, E. M. (2017). *Continuing Bonds in Bereavement: New Directions for Research and Practice*. Routledge.

Krishnan, R., Rajpurkar, P., & Topol, E. J. (2022). Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, *6*(12), 1346–1352.

Krueger, J., & Maiese, M. (2018). Mental institutions, habits of mind, and an extended approach to autism. *Thaumàzein*, *6*, 10–41.

Krueger, J., & Osler, L. (2022). Communing with the Dead Online: Chatbots, Grief, and Continuing Bonds. *Journal of Consciousness Studies*, *29*(9–10), 222–252.

Krueger, J., & Osler, L. (2019). Engineering Affect: Emotion Regulation, the Internet, and the Techno-Social Niche. *Philosophical Topics*, *47*(2), 205–231.

Lemoine, B. (2023, February 27). *“I Worked on Google’s AI. My Fears Are Coming True.”* Newsweek. https://www.newsweek.com/google-ai-blake-lemoine-bing-chatbot-sentient-1783340

Mallory, F. (2023). Fictionalism about Chatbots. *Ergo (Ann Arbor, Mich.)*, *10*(0). https://doi.org/10.3998/ergo.4668

Manmadhan, S., & Kovoor, B. C. (2020). Visual question answering: a state-of-the-art review. *Artificial Intelligence Review*, *53*(8), 5705–5745.

Nguyen, C. T. (2022). Trust as an Unquestioning Attitude. In *Oxford Studies in Epistemology Volume 7* (pp. 214–244). Oxford University PressOxford.

Nguyen, C. T. (2019). Games and the Art of Agency. *The Philosophical Review*, *128*(4), 423–462.

Rani, V., Nabi, S. T., Kumar, M., Mittal, A., & Kumar, K. (2023). Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering. State of the Art Reviews*, *30*(4), 2761–2775.

Reardon, S. (2023, June 14). *AI Chatbots Could Help Provide Therapy, but Caution Is Needed*. Scientific American. https://www.scientificamerican.com/article/ai-chatbots-could-help-provide-therapy-but-caution-is-needed/

Roberts, T., & Krueger, J. (2022). Musical agency and collaboration in the digital age. In K. Bicknell & J. Sutton (Eds.), *Collaborative Embodied Performance: Ecologies of Skill* (pp. 125–140). Bloomsbury Publishing.
Robson, J., & Meskin, A. (2016). Video games as self-involving interactive fictions. *Journal of Aesthetics and Art Criticism*, *74*(2), 165–177.

Rodogno, R. (2016). Social robots, fiction, and sentimentality. *Ethics and Information Technology*, *18*(4), 257–268.

Ruane, E., Birhane, A., & Ventresque, A. (2019). *Conversational AI: Social and ethical considerations*. AICS. http://ceur-ws.org/Vol-2563/aics_12.pdf

Rucińska, Z. (2018). The role of affordances in pretend play. In C. Durt, T. Fuch, & C. Tewes (Eds.), *Embodiment, enaction, and culture: Investigating the constitution of the shared world* (pp. 257–278). MIT Press.

Smuts, A. (2009). What is interactivity? *Journal of Aesthetic Education*, *43*(4), 53–73.

Sweeney, P. (2021). A fictional dualism model of social robots. *Ethics and Information Technology*, *23*(3), 465–472.

Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction*, *11*(4), 1–51.

Tong, A. (2023, March 25). AI company restores erotic role play after backlash from users 'married' to their bots. *The Sydney Morning Herald*. https://www.smh.com.au/world/north-america/ai-company-restores-erotic-roleplay-after-backlash-from-users-married-to-their-bots-20230326-p5cvao.html

Van De Mosselaer, N. (2020). Imaginative desires and interactive fiction: On wanting to shoot fictional zombies. *The British Journal of Aesthetics*, *60*(3), 241–251.

Wiener, A. (2023, November 13). Holly Herndon's Infinite Art. *The New Yorker*. https://www.newyorker.com/magazine/2023/11/20/holly-herndons-infinite-art

Wildman, N. & Woodward, R. (2018). Interactivity, Fictionality, and Incompleteness. In Robson, J. & Tavinor, G. *The Aesthetics of Videogames*. New York: Routledge.

Wittkower, D. E. (2022). What is it like to be a bot? In S. Vallor (Ed.), *The Oxford Handbook of Philosophy of Technology* (pp. 357–373). Oxford University Press.