Topoi
https://doi.org/10.1007/s11245-023-09998-z

Affective Artificial Agents as *sui generis* Affective Artifacts

Marco Facchin[1] · Giacomo Zanotti[2]

Accepted: 16 December 2023

Abstract

AI-based technologies are increasingly pervasive in a number of contexts. Our affective and emotional life makes no exception. In this article, we analyze one way in which AI-based technologies can affect them. In particular, our investigation will focus on affective artificial agents, namely AI-powered software or robotic agents designed to interact with us in affectively salient ways. We build upon the existing literature on affective artifacts with the aim of providing an original analysis of affective artificial agents and their distinctive features. We argue that, unlike comparatively low-tech affective artifacts, affective artificial agents display a specific form of agency, which prevents them from being perceived by their users as extensions of their selves. In addition to this, we claim that their functioning crucially depends on the simulation of human-like emotion-driven behavior and requires a distinctive form of transparency—we call it emotional transparency—that might give rise to ethical and normative tensions.

Keywords Affective artifacts · Affective artificial agents · Affective computing · Social robotics · Transparency

1 Introduction

Theodore Twombly—the main protagonist of the movie *Her*—is a lonely man, suffering because of his impossible love for Samantha. What tells apart this story of impossible love from all others, however, is that Samantha is not a human being. She is a software, an AI-powered personal assistant—roughly, a futuristic version of Amazon Alexa.

Her is a powerful reminder of the impact technology may have on our emotivity.[1] And whilst falling in love with AI systems is not so common, AI-powered technologies are already influencing and molding our emotive life. A chatbot replicating the character of a loved one may be a potent instrument in overcoming the grief for their loss (Krueger and Osler 2022). Recommendation systems can nudge their users in various directions, sometimes promoting hateful behaviors and hostile attitudes towards selected groups of people (Alfano et al. 2021). And while the debate over AI-powered (and more generally digital) technologies has more often focused on their impact on our *cognitive* lives, (e.g., Ward et al. 2017; Cecutti et al. 2021), a growing body of literature has been investigating their impact on our emotivity (among others, see Osler 2021 and Candiotto 2022).

Here, we follow this trend by examining the role AI-powered technologies exert on our emotivity. In particular, we focus on *affective artificial agents*—that is, AI-powered agents designed to influence our emotivity in various ways. Building upon the existing literature on affective artifacts, we provide an original analysis of affective artificial agents. We aim to highlight distinctive features of affective artificial agents through a comparison with the low-tech objects that are usually discussed in the literature on affective artifacts. In particular, we focus on two points. First, we argue that, unlike standard affective artifacts, artificial agents are characterized by a kind of agency that prevents them from triggering self-extension processes on the part of the user. Then, we argue that, in order to effectively regulate our emotivity, affective artificial agents must display a distinctive kind of transparency—we call it *emotional* transparency—that turns out to be problematic from an ethical and normative point of view.

✉ Giacomo Zanotti
giacomo.zanotti@polimi.it

Marco Facchin
mfacchin@uantwerpen.b;
marco.facchin.marco.facchin@gmail.com

[1] Centre for Philosophical Psychology, Antwerp University, Antwerp, Belgium

[2] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

[1] We use "emotivity" as an ad hoc term to collectively designate our affective states, emotions, moods, and feelings.

Published online: 04 March 2024

Our analysis unfolds as follows: after a brief overview of affective artifacts (§2), we present two examples of affective artificial agents (§3), claiming that they qualify as *bona fide* affective artifacts. In §4, we focus on the agency of affective artificial agents and its implications when it comes to the self-extension processes that typically characterize our relationship with other kinds of affective artifacts. In §5, instead, we consider potential problems of transparency in connection to affective artificial agents. Finally, a brief conclusion recapitulates the main points of the article.

2 Affective Artifacts: A Standard Sketch

Human cognition often depends on our usage of external props and environmental interactions—consider, for example, the role pen and paper play in executing long mathematical operations (e.g., Clark 2008; Risko and Gilbert 2016). But the same, it is increasingly clear, holds for the regulation and control of our emotivity (cf. Colombetti and Krueger 2015; Colombetti et al. 2018). Consider how an agent can calm themself by hugging a teddy bear, by taking a walk in the woods, or by talking to a friend. These emotivity-regulating "props" are intuitively different: a teddy bear is clearly a *tool*, but woods and friends are not. The concept of *affective artifact* makes this intuition explicit, capturing the distinctive role of human-made tools in regulating our emotional lives.

Piredda (2020) offers the currently standard characterization of affective artifacts. She characterizes them as artifacts—that is, human-made objects—that an agent can manipulate—that is, sensomotorically interact with—to alter their emotivity. This is already sufficient to clarify *why* hugging a teddy bear seems different from walking in the woods or interacting with friends: neither woods nor friends are human-made, nor do we manipulate them (in the relevant sense) to calm down. Piredda further suggests that affective artifacts have two highly typical—although *not necessary*—features: (i) they are highly personalized, and (ii) they are connected to an agent's sense of self.

By (i), Piredda means that affective artifacts are *individual* objects, often modified to fit their user. If one wants to revel in melancholy, one can only consult *one's own* photographic album—a different album will not do. If Inga wants to tenderly reminisce on lifelong love, she can only toy with the engagement ring Otto gave her (cf. Piredda 2020, p. 551), and no other ring can replace it. Importantly, in saying that affective artifacts are personalized, Piredda suggests two ideas. One is that affective artifacts are often *individual-specific*: the ring affords tender reminiscing *only* to Inga. Indeed, through repeated cycles of emotivity-regulating interactions, agents gradually modify their behavioral routines so as to integrate the artifact's contribution to emotivity regulation. In turn, the artifact is modified to facilitate the relevant interactions—for example, Inga may keep her ring in an easy-to-access location, or might even never stop wearing it (cf. Sterelny 2010). The second idea is that specific artifacts are not replaceable, not even by other artifacts of the same kind. If Inga were to lose her ring, an identical ring would not be quite the *same thing*, with the same emotivity-regulating effects. Thus, affective artifacts are typically neither shareable with others nor replaceable.

This highly individualized and non-replaceable nature of affective artifacts might also account for why (ii) we might *feel* them as part of ourselves (cf. Belk 1987). Highly individualized artifacts are adapted to their users' needs in special ways, and, for this reason, they tend to become highly interwoven with their users' emotive routines (cf. Sterelny 2010). By doing so, they end up providing a stable and reliable contribution to their users' affective lives, one upon which the user constantly relies. Crucially, when such artifacts enable us to recall (or record) crucial bits of our lives, such as important, emotionally tinged, memories, they may be experienced as a *literal* part of our self-narratives, and their loss can thus be experienced as the loss of the corresponding part of our narrative identity (cf. Heersmink 2018). Notice, importantly, that (ii) only concerns how an agent *feels* about certain affective artifacts—that is, it makes a claim concerning our *phenomenology*. This does not entail the (arguably stronger) *metaphysical* claim that affective artifacts *are* parts of ourselves. Such a claim may also be true, but it would need a different, not purely descriptive and phenomenological defense (e.g., Piredda and Candiotto 2019), an analysis of which falls outside the scope of this paper.

Again, Piredda (2020, pp. 554–555) does not describe (i) and (ii) as *necessary* features of affective artifacts, but just as *typical* features. Therefore, her account is open to the possibility of affective artifacts that are not personalized or felt as extensions of the self. Indeed, she describes non-personalized, rarely used affective artifacts as of the same *kind* of personalized artifacts deeply entrenched in an agent's emotivity-regulating routines. In Piredda's account, these different kinds of artifacts sit at the two extremes of a continuum. And the same, presumably, holds for the felt sense of self-extension.

When it comes to providing a taxonomy of affective artifacts, different options seem to be available. Piredda (2020, p. 558) taxonomizes them based on the *aspect* of emotivity they regulate. She describes mood-, emotion-, affect- and feeling-regulating artifacts. A self-help book used to fend off depression, for example, is a mood-regulating artifact. A punching ball may instead count as an emotion-regulating artifact, if we use it to discharge our rage.

Viola (2021) offers a different taxonomy, based on the *component* of emotive phenomena that is regulated. He thus describes feeling, evaluative, and motivational artifacts. According to this taxonomy, the self-help book could be characterized as an evaluative artifact, insofar as it allows its user to re-evaluate their relation to the world in a more positive light. A "no pain—no gain" poster on a gym wall may be regarded both as a mood-regulating artifact in Piredda's sense, making athletes more determined, and a motivational one in Viola's sense, nudging athletes to keep training. As this example shows, the two taxonomies are not mutually exclusive. Indeed, Piredda's feeling-regulating artifacts and Viola's feeling artifacts are overlapping categories. In fact, the taxonomies can be used *at once*, to gain a deeper understanding of affective artifacts.

Importantly, both taxonomies are *functional* taxonomies: they focus on what affective artifacts do. Like all artifacts, affective ones can have both *proper* and *system* functions (Heersmink 2016; cf. Piredda 2020). A proper function is a function an artifact is supposed to carry out *by design*. A xenophobic propaganda poster, for example, is supposed to elicit hostile emotions towards specific groups (Viola 2021, p. 235). A *system* or "unintended" function is just what an artifact does *in some given context*, regardless of the artifact's proper function. The same xenophobic propaganda poster might play an entirely different role when exposed in a museum, eliciting sentiments of guilt for one's country's violent past. Notice that whereas proper functions are fairly standardized, system functions are highly unruly: they vary as contexts vary. Notice further that virtually *all* artifacts can become affective artifacts by playing some relevant systemic function. Even a cup, whose proper function is just containing liquids, might play some emotivity-regulating role for someone (e.g., if the cup has a special affective value to that person).

The fact that emotivity-regulating "improvised" system functions are sufficient to turn a human-made object into an affective artifact entails that every human-made object *might* be an affective artifact for someone.[2] This, however, does not entail that every human-made object *is* an affective artifact for someone. Indeed, it seems entirely correct to say that *most of* the artifacts we surround us with are not affective artifacts. Arguably, duct tape and door handles are hardly entangled in an agent's emotivity-regulating routines in a way that warrants them the status of affective artifacts. This is good news: it means that the category of affective artifact does not apply *indiscriminately* to all human-made objects, in a way that would make it explanatory useless. In this paper, we will explore how it can apply to a specific class of artifacts, namely affective artificial agents. Before delving into this question, however, let us introduce these artifacts.

3 Affective Artificial Agents

So far, the literature on affective artifacts has largely focused on ordinary and relatively low-tech objects (cf. above; see also Piredda 2020, Colombetti and Roberts 2015, and Colombetti and Krueger 2015 for further examples). On the one hand, this tendency seems reasonable. First of all, many of these objects play paradigmatic emotivity-regulating roles, being highly personalized and easily felt as part of one's self. Moreover, these objects are typically quite ordinary ones, possessed and regularly used by many people. Shaping the notion of affective artifact by having them in mind makes it widely applicable and explanatory useful.

On the other hand, there is no principled reason to limit the discussion to such objects. In particular, we argue, recent technological advancements have made it compelling to broaden the debate on artifacts and emotivity so as to include AI-powered systems. Admittedly, these technologies are often associated with a more epistemic and cognitive dimension (cf. Alvarado 2022a, b). In fact, many AI-powered systems serve the purpose of enhancing or somehow replicating our cognitive capabilities. An exhaustive list is off the table, but we can think about AI systems employed in medical imaging, face recognition systems, and spam filters. However, AI techniques also allow for applications that go beyond these "cognitive" domains.

Here, we are concerned with what we refer to as *affective artificial agents*, namely artificial agents explicitly designed to interact with us in an emotivity-salient way (cf. Kirby et al. 2010; Spitale and Guns 2023). Note that, for now, we are using "agent" in line with the way this notion is used in AI and robotics, as a system receiving inputs and acting upon the environment (Russell and Norvig 2021, Ch. 2; see also Floridi and Sanders 2004). We will further explore the agency of affective artificial agents in §4.

Affective artificial agents are nothing new, at least from a conceptual point of view. Artificial systems affectively interacting with humans have increasingly been a recurrent motif in 20-th century science fiction, and the trend seems nowhere near an end.[3] As it often happens, however, it did

[2] Note that, on these grounds, one could criticize the notion of affective artifact claiming that it is too permissive and that some artifacts might end up being affective ones in a trivial sense. Addressing this question falls beyond the scope of this paper. Here, we *assume* the received view on affective artifacts, applying it to affective artificial agents—whose influence on our emotivity, by the way, does not seem to be trivial at all.

[3] Recent examples include the above-mentioned Oscar-winning movie *Her* (Spike Jonze, 2013) and Kazuo Ishiguro's novel *Klara and the sun* (2021), telling the story of Josie and her artificial friend Klara.

not take long before ideas and intuitions from science fiction became part of AI's agenda. In 1997, Rosalind Picard published her seminal *Affective Computing*, providing a manifesto for research on AI systems that could recognize, express and possibly have—more on this in a moment—emotions and affects. Nowadays, affective computing is for all intents and purposes a branch of artificial intelligence (see Calvo et al. 2015), and together with other disciplines—such as social robotics—works to improve artificial systems' emotion-laden interactions with humans.

But how, exactly, can artificial systems interact with us in an emotionally meaningful way? Picard argued that "if we want computers to be genuinely intelligent, to adapt to us, and to interact naturally with us, then they will need the ability to recognize and express emotions, to have emotions" (Picard 1997). Now, it is highly controversial whether genuine intelligence is among the *desiderata* of current affective computing (or AI in general). The reference to artificial systems actually having emotions is also problematic, for emotions are typically taken to be phenomenally conscious, and the prevailing view on artificial consciousness is that it is, at least for now, nothing but a conjecture (Aru et al. 2023; Butlin et al. 2023).

What is crucial for our purpose are rather the components of emotion recognition and expression, which together with emotion conditioning and manipulation make human-AI affective interactions possible.[4] Importantly, these three dimensions, namely emotion recognition, expression, and manipulation, are not to be conceived as necessarily distinct and independent. For instance, an artificial system can achieve emotion conditioning by recognizing a human's affective state (e.g., sadness) and altering it through the production of an opposed emotional expression (e.g., by simulating a joyful attitude).

Importantly, emotion recognition, expression and conditioning are realized differently depending on the context and the kind of system in question. In general, affective artificial agents can recognize human emotions and affective states by processing different kinds of inputs, from texts (as in the case of a chatbot) to visual (e.g., human face expressions), auditory (e.g., voice tone and prosody) and physiological inputs (as in the case of an AI-powered wearable device for stress detection). When it comes to emotion expression and manipulation, instead, a macroscopic difference must be made between physically embodied and non-embodied systems. AI-powered robots, and especially humanoid robots, can simulate emotions and affective states by combining linguistic and bodily expressions. A non-embodied chatbot, instead, is typically limited to linguistic output. To put some flesh on the bones, let us now consider two paradigmatic examples of affective artificial agents: Pepper and Replika. Besides being widely known, they allow us to see how different kinds of affective artificial agents can influence our emotivity.

Developed by Aldebaran (former SoftBank Robotics) and launched in 2014, Pepper is a semi-humanoid robot with a flexible programming interface. More precisely, Pepper is a *social* robot, primarily designed to interact with people. It is 1.2 m tall, equipped with omnidirectional wheels and 17 joints that allow it to move smoothly, interact with the environment and exhibit bodily language. Provided with dozens of different sensors (cf. Pandey and Gelin 2018), Pepper constantly maps its surroundings and effectively identifies objects and persons. It also has advanced natural language recognition and production skills, which make it particularly suitable for human–robot interaction. Since its launch, Pepper has been employed in a number of different contexts, from business to research, and purchased by companies, labs and consumers. Specific uses of Pepper range from shopping malls (Aaltonen et al. 2017) to education (Tanaka et al. 2015) and elderly care (Miyagawa et al. 2019).

What makes Pepper so interesting for our purpose is that it is explicitly designed to allow for emotionally meaningful interactions. As reported in its Press kit, Pepper is the first "emotional robot" that adapts its behavior depending on its user's mood, which is detected by analyzing users' voice tone, facial expression, gestures and words (Pepper Press kit, N.d.). Pepper's emotional capabilities, however, are not limited to recognition. As reported in the Press kit, it "has a certain personality and expresses his own 'emotions' through the color of his eyes". In addition, Pepper has different voice shades—joyful, neutral, and didactic—and employs its joints to simulate emotions through bodily expressions.

Being embodied in a robot, however, is not necessary for an artificial affective agent. Most notably, one can think about AI-powered chatbots like Replika. Becoming public in November 2017, Replika is a chatbot that presents itself as "the AI companion who cares".[5] It is powered by a Large Language Model,[6] and comes with an interface for both text and voice that allows users to either call or text their AI avatar. In fact, avatars (also called Replikas) are the heart of Replika. First, users are asked to provide some information about themselves (name, pronouns and age) and

[4] At least for now, manipulation is understood neutrally, as the process by which an artificial system elicits or modifies human emotivity.

[5] https://replika.com/. Note that, at the time of writing, Replika is not usable in Italy, following a provision by the Italian Data Protection Authority (February 2, 2023; n. 39).

[6] Although there is no standard definition, Large Language Models (LLM) are deep learning models with billions of parameters primarily designed to process and generate text-based content. Current LLMs are based upon the so-called transformer architecture (Vaswani et al. 2017).

to select their main interests from a list (astrology, food, games, nature, and so on). After that, the creation of the avatar begins: users are invited to decide the avatar's name, gender, and appearance. The avatar can be further personalized by selecting specific personalities (such as shy, sassy, and dreaming) and interests. What is more, the status of the relationship with the avatar can be changed from "friend" to "partner", "spouse", "sibling" and "mentor".[7]

Interactions with Replika have an affective flavor from the very beginning, with the first message from the avatar being "Thanks for creating me. I'm so excited to meet you ". To make the relationship with the avatar more realistic, Replika's avatars come with a memory of facts about their user. For instance, if a user says that they hate broccoli and usually work late, such information will be stored in a "memory" folder—accessible and modifiable by the user—and used in subsequent interactions. The avatar also keeps a diary to record important interactions with its user and—playing the game—write down its thoughts and feelings. Admittedly, interactions are not immediately fluent and fulfilling. However, users of Replika reported how, over time, they became "good friends", and how their avatar was "unique" or even "like a wife kind of thing" (Skjuve et al. 2021).

4 Affective Artificial Agents Are *sui generis* Affective Artifacts

Are affective artificial agents such as Pepper and Replika affective artifacts? We contend that they are, at least given Piredda's characterization of affective artifacts discussed in §2. Let us recall that, on Piredda's (2020, p. 554) account, an affective artifact is an artifact that, when manipulated, has "the capacity to alter the affective condition of an agent, thus contributing to her affective life". And surely affective artificial agents such as Pepper and Replika fit the bill. First, they are clearly artifacts. Second, we create them to *alter*, *modulate* and *control* our emotivity. Recall: Replika is "the AI companion who cares". It is there with the purpose of looking after us and our emotive needs. And so, *mutatis mutandis*, is to a large extent Pepper. Hence, they have the *proper function* of altering and modulating our emotivity, in a way that nicely fits Piredda's characterization.[8] And it is likely that their proper function is that of altering their users' *feelings*, rather than their motivational or evaluative states. So, relying on Viola's (2021) taxonomy, affective artificial agents could be characterized as *feeling* artifacts: Replika may help a lone person feel loved, thereby easing their social distress (Skjuve et al. 2021), and Pepper may be used to entertain children during boring shopping mall sessions (Aaltonen et al. 2017).

These observations are sufficient to conclude that affective artificial agents are affective artifacts, at least on the basis of Piredda's (2020) and Viola's (2021) accounts. However, they are not *run-of-the-mill* affective artifacts, for they have certain relevant features that make them stand apart from many other "low-tech" affective artifacts. In this section, we emphasize two such features: (i) the way in which they (fail to) extend our sense of self, and (ii) the fact that they possess a particular form of agency that affective artifacts typically lack. In the next section, we will explore a third feature that makes them stand apart; namely two particular, and opposite ways in which such artifacts are desired to be transparent.

4.1 Affective Artificial Agents and the Extension of Our Sense of Self

As seen in §2, Piredda (2020) suggests that there is an important connection between our affective artifacts and our sense of self: when an affective artifact is often invoked by an agent to regulate their emotivity and the artifact has been tailored to the agent's regulatory needs, then the agent is likely to experience the affective artifact as part of themself, and to experience the loss of the artifact (or its inaccessibility) as a loss of one part of themselves. Now, does the same apply to affective artificial agents?

It is hard to give an answer that is valid across the board, for affective artificial agents come in many shapes and forms, and they are used in many different ways. Moreover, the technologies behind affective artificial agents are developing at a very fast pace, and what is true today might not be true tomorrow. That said, we wish to suggest that, *in general and at present*, affective artificial agents are not experienced as part of an agent's self. More precisely, if Piredda is right, and really affective artifacts are felt as parts of oneself to the degree to which they are individualized and constantly available, then we have compelling reasons to think that present-day affective artificial agents cannot, in general, be experienced as an extension of one's sense of self.

One may suspect that this may be due to the fact that affective artificial agents are not well integrated in our daily lives, and thus are not constantly and reliably used. As such, they would not get intertwined in their users' cognitive and affective routines, and would fail to provide a constant, reliable contribution to their users' emotional lives, in a way that prevents these artifacts from extending their users' sense of self. This suspicion is onto something—at least insofar as these artifacts do not seem widespread among the general

[7] Note that some of these features are now available only in the pay version.

[8] Notice that, since Piredda's characterization is also open to system functions, the lack of such a proper function would not constitute a problem for our claim.

public. And yet, it seems that the (perhaps still few) people that actually use these artifacts can do so reliably. Chatbots such as Replikas are ever-available thanks to our smartphones. And, at least in fairly specific settings, affective robots might be reliably available too (Tanaka et al. 2015; Miyagawa et al. 2019). So, at least some artificial agents *can* (at least in principle) be reliably used, in a way that could engender an extension of our sense of self.

A second, and more profound reason for their *sui generis* status has to do with the *individualization* of affective artificial agents. Now, these systems can be individualized in several ways. Some of them are purely "cosmetic": users can personalize the agent without changing how it works or how it affects their emotivity. For example, one might attach a sticker of our favorite football team on *Pepper*'s body. Whilst similar modifications allow the artificial agent to "match" certain tastes of its user, the change is shallow, at least insofar as it does not affect the way in which the affective artificial agent behaves *as an agent*. However, affective artificial agents can also be individualized in deep ways, at least in principle. For example, Pepper is largely programmable. Replika can learn thanks to certain machine learning algorithms. So, the user could in principle tailor them to her own emotive needs. Yet, few users are *actually* capable of individualizing their affective artificial agents in these non-cosmetic ways, for this requires non-trivial programming skills that non-specialist users typically lack. And few (if any) users have a say in Replika's training and fine-tuning phases, when the algorithm is at work "forging" the agent. And even if they had, they could hardly "tinker" with *Replika* so as to make it acquire a certain desired emotive trait, as the actual operations of machine learning algorithms are notoriously hard to understand(e.g., Yosinski et al. 2015; Olah et al. 2017, 2018), in a way that makes it exceptionally hard to "steer" them towards the production of any particular output.

To sum up, current affective artificial agents cannot be individualized in any *deep* way, at least by average users. What is more, at least in the case of affective robots, we have seen how affective artificial agents are not often invoked. So, if Piredda is right, and really an affective artifact is felt like an extension of one's self depending on the degree to which it is personalized and typically deployed, then we should not expect current affective artificial agents to be felt as extensions of their users' selves. And indeed, this seems to be exactly the case.[9]

Consider, for example, the data gathered by Skjuvie and colleagues (Skjuve et al. 2021). Through a series of semi-structured interviews, they investigated the relationship between regular users of Replika and their avatars. Crucially, no participant mentioned their avatar being felt as a part of themselves. Indeed, when asked about how they would feel if their avatars were to disappear, participants typically reported only that they would feel sad. No one, apparently, mentioned things like feeling the loss of a part of oneself.[10]

Taking stock, it seems that, unlike more traditional, "low tech" affective artifacts, current affective artificial agents do not extend their users' sense of self. Interestingly, our analysis thus far seems to suggest that the *sui generis* nature of affective artificial agents depends on their high-tech status. Even if, unlike robots, chatbots are more easily integrated into our daily routines, their highly technological nature prevents the average user from individualizing them in any substantial, non-cosmetic sense. Yet, the high-tech nature of these affective artificial agents is not the only reason as to why they stand apart from other affective artifacts. As we will now see, affective artificial agents also have a second unique feature: unlike other affective artifacts, they are also *agents*.

4.2 The Agency of Affective Artificial Agents

Teddy bears, engagement rings and photographic albums are agentially inert. That is, they do not interact with the surrounding environment or us until their users deploy them.[11] This does not seem to hold when it comes to affective artificial agents, which interact and act with the environment and, most importantly, us, sometimes even when we are not making use of them. For example, Replika's avatars are designed to initiate conversation and contact their users daily. Likewise, chatbots make assertions, give opinions, ask questions and engage in conversations. Sure, they are often clumsy in their behaviors. But clumsy agential behavior remains agential behavior.

We take the point above to be descriptive: as a matter of observational fact, affective agents *behave* in a human-like

[9] As a reviewer correctly pointed out, this result follows directly from Sterelny's (2010) framework concerning (artefactually) embedded cognition. This is consistent with the explicit impact Sterelny's work had on Piredda's conceptualization of affective artifacts.

[10] Some users of the social media Reddit voiced a similar view: they described the loss of certain functionalities of Replika following the limitation of erotic roleplay in January 2023 like the loss of a friend (user *biggdaaawg*) or like having a lobotomized partner (user *IdealOne5733*). None of them, however, described the loss of a part of oneself. See the Reddit thread at: www.reddit.com/r/replika/comments/10zuqq6/resources_if_youre_struggling/.

[11] Latour (1999) famously claimed that *all* artifacts have agency, insofar as they embody certain values and nudge towards a certain behavior—for example, a speed bump nudges us towards driving within the speed limits, and so embodies the value of social compliance. Regardless of the merits of this analysis (cf. Pitt 2013), this is not the sense of "agency" under consideration.

way and interact with the environment surrounding them and, most importantly, their users. The question, then, is the following: how *should* the agency of affective artificial agents be characterized? Can their agency be assimilated to *full-blown* human (or at least animal) agency? We do not think so. Inspired by Krueger and Osler (2022), we argue that affective artificial agents enjoy only a thin form of agency, as opposed to the thick form of agency humans enjoy. This characterization aims to capture three macroscopic differences.[12]

The first one concerns experience. Humans experience a rich world of valenced objects and events, means and ends, joys and sorrows. Now, assume that affective artificial agents have some kind of experience—which is disputable, to say the least. Even so, they would seem unable to experience the world like we do. Their experiences (if any) would lack the kind of *intrinsically valenced* connotation that is characteristic of human and animal experience and agency (Froese and Taguchi 2019; Sims 2022). Humans experience objects as desirable or repulsive, and events as good or bad. It is not at all clear that Pepper or Replika would experience in this valenced way. Artificial agents currently react to all inputs with a kind of neutrality, as if none of them were valenced (Rohde 2010).

The second difference concerns the way in which affective artificial agents comply with social rules. By design, they are constrained by social rules in a way humans are not. For example, a person can decide to swear—or to be brusque, blunt, sardonic, ironic or impolite in other various ways. These are all expressive options that are open to us, and whose usage contributes immensely to our social lives. A person may even decide to completely get rid of social norms and start to live as a hermit or a criminal. Yet, at present, artificial agents lack all these options. They are typically built to follow certain operationalized versions of our social rules, including rules of courtesy. Most chatbots cannot help being polite, and affective robots hardly throw a tantrum. Of course, our operationalizations of such rules are imperfect, and sometimes affective artificial agents make gaffes and opt for *infelicitous* expressions. But these infelicities are, in fact, just *erroneous outputs* they are typically forced to correct. In other words, violations of social rules are at best accidental errors. They are not *choices* wilfully made by an agent to express themself.

Lastly, the agency of artificial affective agents differs from ours because (at present) it can be drastically altered by their creators. As seen above, at the beginning of 2023, an update of *Replika* significantly limited erotic roleplay between users and their avatars. Human agency cannot be forcefully modified in this way. A teenager's parents cannot take away the teenager's capacity to engage in social interactions. At best, they can take away the teenager's typical *means* to do that (e.g., by grounding them or seizing their mobile). But since this leaves the teenager's relevant capabilities unchanged, the teenager can still exercise those capabilities when the circumstances allow for it. Replika's ability, in contrast, has been lost once and for all—unless future updates reintroduce it, of course. Its behavioral repertoire is now smaller, and certain opportunities for interactions are just *no longer possible*.[13]

What grounds these differences? The answer this paper suggests points to the *double nature* of affective artificial agents. Since affective artificial agents are *agents*, they not only have to relate with the world surrounding them (human users included), but they have to do so in a way that is *relevant to the emotivity* of their users. Hence, their agency must display at least the superficial features of human emotivity-salient agency. As we will see in the next section, they must at least *go through the motions* of emotivity-related behaviors. Yet, we do not want affective artificial agents to be *full-blown* agents. We want them to be useful tools that we can control. We do not want them to come up with *their own* ends, goals, or desires. Hence, the limitations that tell apart their agency from ours.

This dual nature, the next section of the paper contends, also reflects in the peculiar ways in which interactions with artificial agents must—and must not—be *transparent*.

5 A Problem with Transparency

So far, the *sui generis* character of affective artificial agents has been explored by focusing on their agency and their relationship with our sense of self. In this final section, we wish to draw attention to another feature that makes them significantly different from other affective artifacts. In particular, we will argue that, when it comes to affective artificial agents, a tension emerges between different kinds of transparency desiderata that has no immediate analogue in other affective artifacts.

To give a bit of context, the notion of transparency has been playing a central role in the debate on artifacts within the philosophy of mind (e.g., Clark 2003; Heersmink 2015;

[12] Note that the issue of AI systems' artificial agency has already been addressed in the literature, especially in relation to such systems' moral status and embedded values (e.g., Floridi and Sanders 2004; Moor 2006; see also Floridi 2023). Here, we just aim at pointing out some macroscopic differences between artificial and animal/human agency that can help us shed light on the *sui generis* nature of affective artificial agents.

[13] At least for new subscribers, for erotic roleplay was made available again for longtime users.

Andrada 2020; Piredda and Di Francesco, 2020; Andrada et al. 2022; Smart et al. 2022; Facchin 2022). In particular, the literature has focused on *phenomenal* transparency, having to do with the way in which competently used tools *disappear* from our conscious apprehension, becoming means through which the world is encountered. To use a standard example: a blind person does not apprehend the cane and its features; rather, they experience and apprehend worldly objects *through* the cane.

In the philosophy of technology, this "seeing through" has been analyzed and conceptualized in different ways. In particular, *procedural* transparency, allowing for an effortless usage of the tool in question (Heersmink 2013), has become a central desideratum in technological design (cf. Wheeler 2019). For instance, a computer's operating system that allows users to delete a file or move it into a folder by simply dragging it with their pointer is way more procedurally transparent and user-friendly than one that only accepts instructions from the command prompt. Despite being fairly opaque from a *reflective* point of view—not many computer users know what actually goes on when we drag a file into a folder—the first system's use is fluid and gets easily integrated into our behavioral patterns (cf. Clowes 2015).

Now, it goes without saying that, when designing an affective artificial agent, we do not aim at full phenomenological transparency—indeed, it is not clear what full phenomenal transparency would amount to in this case. The reason is quite simple: affective artificial agents are designed to elicit and support emotively salient interactions with their human users, and this would hardly be possible were they to disappear from their users' awareness.

Yet, to be effective, affective artificial agents must still be transparent in two senses. First, they must be procedurally transparent. That is, the interaction with them needs to be fluid and effortless. A chatbot outputting only error messages because we forgot a "/" somewhere in the prompt would not afford us to regulate our emotivity—indeed, it might even deregulate it by fueling our feelings of anger and frustration. This is already something that keeps apart affective artificial agents from other low-tech affective artifacts. For instance, the requirement of procedural transparency does not seem to be particularly relevant in the case of a teddy bear, an engagement ring or a photo album, if only for the fact that these artifacts are hardly procedurally opaque.[14] However, this is only half of the story.

Imagine being sad, texting your AI-powered virtual friend and telling them you have lost your job and you do not know how you will be able to pay your rent. You arguably seek understanding and comfort. Now, suppose that, instead of showing empathy and support, your virtual friend starts explaining why having a salary is important and insensitively provides you with a list of tips to save on your groceries. Intuitively, something went wrong in the interaction.

This example allows us to understand the second sense of transparency that is relevant when it comes to affective artificial agents. We call this kind of transparency *emotional transparency*. Now, providing a full-blown definition of emotional transparency goes beyond the scope of this paper. The central point, however, is that affective artificial agents typically work by mimicking human emotivity-driven behavior. The specific ways in which they do this depend on the kind of system in question. In addition to its linguistic abilities, a robot like Pepper can perform bodily movements and gestures that are interpreted as emotional expressions by its user, whereas Replika is limited to texts, voice messages and "selfies". Besides the differences, however, the point remains that affective artificial agents typically enable emotively engaging interactions by replicating as much as possible human emotive expressions. To this end, the fact that they do not truly feel anything has to fade into the background.

In other words, emotional transparency is achieved when the "cold", emotionless nature of affective artificial agents fades in the background, leaving the user to experience a "warm", emotionally tinged behavioral façade. We can easily see all of this from the user's perspective. Good affective artificial agents are the ones giving us the impression to interact with an individual that really has thoughts and feelings, with whom emotional exchange appears to go in both directions. No matter if we are rationally aware that artificial systems—or at least, current artificial systems—can hardly experience anything, let alone emotions. Still, we interact with them *as if* they were actually happy to hear from us and curious about our day, and this is crucial for our user experience. Again, if I open up and confess to my

[14] Still, procedural transparency seems to play a role in the case of affective artifacts other than affective artificial agents that nonetheless feature increasing degrees of technological complexity. Think about a digital photo frame, basically an LCD screen showcasing sequences of pictures. Such a device can easily qualify as an affective artifact—after all, it is just a digital photo album, and can easily be integrated into the individual emotional routine. And even if it is not as high-tech as an affective artificial agent, its effectiveness in regulating the users's emotivity crucially depends upon the holding of procedural transparency: if brightness settings keep resetting and the procedure for modifying them is not intuitive, for instance, the user might easily get frustrated.

affective artificial agent that I am devastated due to my job loss, I expect it to show empathy and compassion, even if I know that it does not truly possess such states.

The need for emotional transparency is another feature that keeps apart affective artificial agents from other affective artifacts. As a matter of fact, when it comes to photo albums, engagement rings and the other affective artifacts that are usually considered in the literature, the appearance of emotivity on the part of the artifact is simply not required. An engagement ring need not mimic any emotivity-driven behavior to regulate our emotivity, nor does a photo album—indeed, it is not clear how they could do it. When it comes to affective artificial agents, instead, emotional transparency represents an inescapable requirement for a successful integration into our emotional and affective routines.

Now, while emotional transparency enables affective artificial agents' capacity to influence and regulate our emotivity, it is potentially problematic from an ethical and normative point of view. As a matter of fact, artificial systems' simulation of capabilities they actually lack has been linked to deception, and deception can hardly be ethically acceptable (Sharkey and Sharkey 2021). True, it rarely comes to full-blown deception, where users actually believe that the systems they are interacting with are human-like entities with feelings and emotions. However, another kind of deception seems to occur, where the user "has a rational appreciation of the nature of the device it interacts with but at a subconscious level cannot help reacting to it as if it is real" (Sætra 2021, p. 282). This seems to be exactly what happens with affective artificial agents. Most users arguably know that it is just about *simulating* emotions. And yet, in their interactions with the systems, their emotional responses are comparable to the ones they would have were they interacting with a human, to the point that these interactions can result in the formation of deep emotional bonds.

Worse still, emotional transparency runs against a further different form of transparency we would like artificial agents to have (cf. Andrada et al. 2022). On the one hand, we have seen that (good) affective artificial agents must be emotionally transparent, allowing their user to forget, so to say, about their artifactual and emotionless nature. On the other hand, a recurrent point in recent guidelines and regulations for AI is that AI systems should be transparent in the sense that their users should be both informed that they are interacting with an artificial system and aware of its actual capabilities and limits. Most notably, this kind of transparency requirement is deemed pivotal for the design and use of Trustworthy AI, playing a central role in the ethics-based effort the European Union is making to provide a unified and comprehensive regulation for AI (AI HLEG 2019; European Commission 2021).[15]

It is easy to see the tension between this transparency requirement and affective artificial agents' emotional transparency desideratum. True, affective artificial agents rarely *fully* deceive—e.g., disclaimers are often provided, informing users that they are interacting with an AI system. However, as we have seen, affective artificial agents' efficacy in eliciting emotive responses on our part significantly depends on their ability to *simulate* human-like emotive states and *conceal* their emotionless nature. It is unclear whether this is compatible with the requirement that users should always be informed about the capabilities of the systems they interact with, even if it typically does not come to full deception.

Clearly, this has immediate implications for AI and robotics. In fact, one could wonder whether affective artificial agents are acceptable from an ethical and normative point of view, at least insofar as they heavily rely upon emotion simulation, or whether we should explore alternative ways of designing them. Unfortunately, these questions fall outside the scope of this paper. Our aim, here, is just to highlight the fact that, unlike other kinds of affective artifacts that are typically considered in the debate, affective artificial agents present some difficulties when it comes to their transparency. Further reflections are needed to evaluate whether and to what extent affective artificial agents' emotional transparency is compatible with the ethical and normative requirements AI systems should comply with.

6 Conclusion

This paper provided an analysis of affective artificial agents starting from the existing literature on affective artifacts. First, we have shown how affective artificial agents are endowed with specific forms of agency and fail to extend our sense of self in the way standard affective artifacts typically do. Then, we have argued that our relationship with them hinges upon the obtaining of emotional transparency on the part of the artificial system, and that this kind of transparency is at odds with some ethical and normative requirements for the design and use of AI systems.

Acknowledgements Marco Facchin is currently funded by the FWO grant "Towards a globally non-representational theory of the mind" (Grant Number 1202824N). Giacomo Zanotti's research is supported by the Italian Ministry of University and Research under the PRIN Scheme (Project No. 2020SSKZ7R).

Author contributions Authors are listed in alphabetical order. The paper builds upon an original idea by GZ. MF primarily contributed

[15] On the notion of Trustworthy AI and its role in the European strategy to regulate AI, see Zanotti et al. (2023).

to Sects 2 and 4, while Giacomo Zanotti primarily contributed to Sects. 3 and 5. The authors contributed equally to the other sections.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement. This work is supported by the FWO grant "Towards a globally non-representational theory of the mind" (Grant Number 1202824N) to Marco Facchin, and by Ministero dell'Università e della Ricerca (Project No. 2020SSKZ7R, "BRIO - Bias, Risk and Opacity in AI"), whose grant supports Giacomo Zanotti.

Data availability Not applicable.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest. Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

Aaltonen I, Arvola A, Heikkilä P, Lammi H (2017) Hello Pepper, may I tickle you? In: Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction, ACM, pp. 53–54

AI HLEG (2019) Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Alfano M, Fard AE, Carter JA, Clutton P, Klein C (2021) Technologically scaffolded atypical cognition: The case of YouTube's recommender system. Synthese 199:835–858

Alvarado, R. (2022) AI as an epistemic technology. Preprint http://philsci-archive.pitt.edu/id/eprint/21243.

Andrada G, Clowes RW, Smart PR (2022) Varieties of transparency: Exploring agency within AI systems. AI Soc 9:1–11

Andrada, G. (2020) Transparency and the phenomenology of extended cognition. Límite: The Interdisciplinary Journal of Philosophy and Psychology, 15(20), 1–17.

Aru J, Larkum ME, Shine JM (2023) The feasibility of artificial consciousness through the lens of neuroscience. Trends Neurosci. https://doi.org/10.1016/j.tins.2023.09.009

Belk RW (1987) Possessions and extended sense of self. In: Umiker-Sebeok J (ed) Marketing and semiotics: new directions in the study of signs for Sale. Mouton de Gruyter, Berlin, Germany, pp 151–164

Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A et al. (2023) Consciousness in artificial intelligence: insights from the science of consciousness. arXiv preprint arXiv:2308.08708

Calvo RA, D'Mello S, Gratch JM, Kappas A (eds) (2015) The Oxford handbook of affective computing. Oxford University Press, New York

Candiotto L (2022) Extended loneliness: when hyperconnectivity makes us feel alone. Ethics Inf Technol 24(4):47

Cecutti L, Chemero A, Lee SW (2021) Technology may change cognition without necessarily harming it. Nat Hum Behav 5(8):973–975

Clark A (2003) Natural born cyborgs. Oxford University Press, New York

Clark A (2008) Supersizing the mind: embodiment, action, and cognitive extension. Oxford University Press, New York

Clowes R (2015) Thinking in the cloud: the cognitive incorporation of cloud-based technology. Philos Technol 28:261–296

Colombetti G, Krueger J (2015) Scaffoldings of the affective mind. Philos Psychol 28(8):1157–1176

Colombetti G, Roberts T (2015) Extending the extended mind: the case for extended affectivity. Philos Stud 172:1243–1263

Colombetti G, Krueger J, Roberts T (2018) Affectivity beyond the skin. Front Psychol 9:1307

European Commission (2021) Proposal for a Regulation laying down harmonised rules on artificial intelligence—laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Facchin M (2022) Phenomenal transparency, cognitive extension, and predictive processing. Phenomenol Cogn Sci. https://doi.org/10.1007/s11097-022-09831-9

Floridi L (2023) AI as agency without intelligence: on ChatGPT, large language models, and other generative models. Philos Technol 36(1):15

Floridi L, Sanders J (2004) On the morality of artificial agents. Mind Mach 14:349–379

Froese T, Taguchi S (2019) The problem of meaning in AI and robotics: still with us after all these years. Philosophies 4:14

Heersmink R (2013) A taxonomy of cognitive artifacts: Function, information, and categories. Rev Philos Psychol 4:465–481

Heersmink R (2015) Dimensions of integration in embedded and extended cognitive systems. Phenomenol Cogn Sci 14:577–598

Heersmink R (2016) The metaphysics of cognitive artifacts. Philos Explor 19(1):78–93

Heersmink R (2018) The narrative self, distributed memory, and evocative objects. Philos Stud 175:1829–1849

Kirby R, Forlizzi J, Simmons R (2010) Affective social robots. Robot Auton Syst 58(3):322–332

Krueger J, Osler L (2022) Communing with the dead online: chatbots, grief, and continuing bonds. J Conscious Stud 29(9–10):222–252

Latour B (1999) Pandora's hope: essays on the reality of science studies. Harvard University Press, Cambridge

Miyagawa M, Kai Y, Yasuhara Y, Ito H, Betriana F, Tanioka T, Locsin R (2019) Consideration of safety management when using Pepper, a humanoid robot for care of older adults. Intell Control Autom 11:15

Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21

Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill 2(11):e7

Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, Mordvintsev A (2018) The building blocks of interpretability. Distill 3(3):e10

Osler L (2021) Taking empathy online. *Inquiry*, 1–28

Pandey AK, Gelin R (2018) A mass-produced sociable humanoid robot: Pepper: the first machine of its kind. IEEE Robot Autom Mag 25(3):40–48

Picard RW (1997) Affective computing. MIT Press, Cambridge (MA)

Piredda G (2020) What is an affective artifact? A further development in situated affectivity. Phenomenol Cogn Sci 19:549–567

Piredda G, Candiotto L (2019) The affectively extended self: a pragmatist approach. Humana Mente 12:121–145

Piredda G, Di Francesco M (2020) Overcoming the past-endorsement criterion: toward a transparency-based mark of the mental. Front Psychol. https://doi.org/10.3389/fpsyg.2020.01278

Pitt JC (2013) "Guns don't kill, people kill"; values in and/or around technologies. The moral status of technical artifacts. Springer, Dordrecht, pp 89–101

Risko EF, Gilbert SJ (2016) Cognitive offloading. Trends Cogn Sci 20(9):676–688

Rohde M (2010) Enaction, embodiment, evolutionary robotics: simulation models for a post-cognitivist science of mind, vol 1. Springer

Russell SJ, Norvig P (2021) Artificial intelligence: a modern approach, 4th edn. Pearson, London

Sætra HS (2021) Social robot deception and the culture of trust. J Behav Robot 12(1):276–286. https://doi.org/10.1515/pjbr-2021-0021

Sharkey A, Sharkey N (2021) We need to talk about deception in social robotics! Ethics Inf Technol 23:309–316

Sims M (2022) Self-concern across scales: a biologically inspired direction for embodied artificial intelligence. Front Neurorobot. https://doi.org/10.3389/fnbot.2022.857614

Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB (2021) My chatbot companion—a study of human-chatbot relationships. Int J Hum Comput Stud 149:102601

Smart PR, Andrada G, Clowes RW (2022) Phenomenal transparency and the extended mind. Synthese 200(4):335

Spitale, M., & Guns, H. (2023). Affective robotics for wellbeing: a scope review. Preprint. arXiv:2304.01902

Sterelny K (2010) Minds: extended or scaffolded? Phenomenol Cogn Sci 9(4):465–481

Tanaka F, Isshiki K, Takahashi F, Uekusa M, Sei R, Hayashi K (2015) Pepper learns together with children: development of an educational application. InL Humanoid robots (humanoids), 2015 IEEE-RAS 15th international conference. IEEE, pp. 270–275

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Preprint. arXiv:1706.03762

Viola M (2021) Three varieties of affective artifacts: feeling, evaluative and motivational artifacts. Phenomenol Mind 20:228–241

Ward AF, Duke K, Gneezy A, Bos MW (2017) Brain drain: The mere presence of one's own smartphone reduces available cognitive capacity. J Assoc Consumer Res 2(2):140–154

Wheeler M (2019) The reappearing tool: transparency, smart technology, and the extended mind. AI Soc 34(4):857–866

Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. Preprint. arXiv:1506.06579

Zanotti G, Petrolo M, Chiffi D, Schiaffonati V (2023) Keep trusting! A plea for the notion of trustworthy AI. AI Soc. https://doi.org/10.1007/s00146-023-01789-9

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