Fictionalism about Agentic AI

Anthony R. J. Fisher[1,*]

[1]Gonzaga University, Department of Philosophy, Robinson House, Spokane, WA, USA

Abstract

This paper motivates, sketches, and argues for a new meta-ethical theory in AI Ethics, namely, fictionalism about AI systems, especially agentic AI. On this view, AI agents are not moral agents and not truly held accountable for their decisions or outputs, but they are according to the fiction. It is also argued that the fiction is useful and that developers and deployers of AI systems should be explicit in setting up the context so that it is clear to users that an AI agent is a fictional individual. Fictionalism has many attractive features; importantly, it allows for liberal licence to build responsible AI frameworks where some notions of responsibility are applied to AI systems directly (within the fiction), which can fill responsibility gaps.

Keywords

fictionalism, agentic AI, moral status, accountability, responsible AI

1. Introduction

With the exponential growth of AI, and in particular generative AI, over the last few years, researchers in academia and industry recognise the many ethical problems with the development and deployment of AI technology, not to mention with use in the market by the public. Philosophers, Computer Scientists, and Industry have zeroed in on the label of 'responsible AI' as best capturing, in a broad sense, ethical approaches to AI.

At the cutting edge of ethical problems in AI are AI agents (hereafter, agentic AI). An AI agent is an AI system that has a high level of autonomy with the authority to generate outputs and execute actions independently. An AI agent can be assigned a task and go off on its own to complete it. Agentic AI has two sources of autonomy: first, its sophisticated design, involving AI models, software, and cloud services, grants it a high level of autonomy, which can reason and learn, do research, and formulate its own methods of inquiry; second, the ecosystem that it is in, coupled with some sort of authority, grants it power, where some of that power is assigned by a designer or user. Agentic AI is of a different category to the two-way, back-and-forth ChatGPT interaction. It is also different to semi-autonomous embedded AI systems such as self-driving cars. It has been stated that agentic AI is the next developmental stage of generative AI. Many workflows are projected to transition to agentic AI [1]. As of now, the technology is constrained by the current ecosystem that AI agents inhabit. But as ecosystems incorporate AI deep into every aspect of all pieces of software (and down into the operation system of computers) AI agents will have far-reaching abilities to do a wide range of things; moreover, with new protocols (such as MCP) AI agents will be able to interact with each other to accomplish tasks, perhaps more efficiently and in novel ways.

For discussion of responsible AI in applied ethics, see [2], [3], [4]. In Industry, some of the main tech companies explicitly sign on to a responsible AI framework. Microsoft has the Office of Responsible AI, which releases annual reports. Google has responsible AI testing units governed by 'AI Principles', one of which is the principle of responsible AI. Like Microsoft, Google releases Responsible AI progress reports (see https://ai.google/responsibility/principles/).

The ethics of agentic AI is in its infancy, but central questions in AI Ethics, especially to do with notions of responsibility, turn on what we say about agentic AI. We need to get clear on the status

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

[*]Corresponding author.

fishera@gonzaga.edu (A. R. J. Fisher)
0000-0001-5351-4521 (A. R. J. Fisher)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-16

of agentic AI before we can get clear on how AI should be held responsible (if at all). After all, it is one clear case where we might intuitively say that AI is responsible in virtue of the fact that the AI system in question is an agent, with some degree of autonomy and without a human in the loop. In addition, the topic should be examined from an ethical lens, because of its projected dominance in AI technologies.

In this paper, I sketch a meta-ethical theory called 'fictionalism about agentic AI'. This theory states that an AI agent should be construed as a fictional individual. Fictionalism about agentic AI addresses the question of whether AI can truly be held accountable for its decisions and covers the topic of whether AI is a moral agent or merely a tool. My answer to both is that 'it depends', and that it depends on the fiction that is constructed around the AI agent by developers and in some contexts by users of AI systems. In what follows, I explain the relevant background features of fictionalism in other areas of philosophy such as metaphysics, moral philosophy, and philosophy of art. I then develop fictionalism about agentic AI and apply fictionalism to some ethical issues concerning accountability (responsibility) and AI as a moral agent versus as a tool. I conclude that fictionalism about agentic AI is useful for creating and sustaining responsible AI frameworks that ensure trustworthy AI.

In my view, the goal of trustworthy AI is not to be reduced to the relational notion of whether users trust AI systems or AI models. Rather, trustworthy AI at its core is about responsible design, deployment and use of AI technologies. One barrier to trustworthy AI is the creation and persistence of responsibility gaps. A responsibility gap is defined in terms of a failure somehow to attribute responsibility to someone, something, some group, etc. For example, suppose an AI agent is assigned a task that it completes autonomously but results in a harmful outcome towards a human person. Who or what is responsible? If no one or nothing is responsible, then a responsibility gap has been created. I predict that without the fiction and an explicit recognition of the fiction, more and more responsibility gaps will open up and persist. But if it is clear and explicit in the fiction that the AI agent is responsible, then something can be held to account; to be sure, some conceptual engineering is required to make sense of the way in which the AI in the fiction is held to account, but that is a distinct matter. Those details aside, which are to be determined by designers and regulators, if there is a sense in which the AI agent is held to account, then certain responsibility gaps can be filled. The fiction might also explicitly state that the AI agent is not responsible, which would make it clear who is, such as the designer, deployer, or user behind the agent.

Before I continue I must state two limitations of this paper. First, it is a short paper, so some points cannot be elaborated upon. Even though I offer a sketch of fictionalism, the paper contains a new meta-ethical perspective worthy of further exploration and discussion. Offering the barebones of this new perspective is the main, theoretical contribution of this paper. Second, fictionalism about agentic AI is articulated primarily as a meta-ethical theory in AI Ethics. The paper does not go into specific detail in relation to empirical studies about how a normative framework can be developed to ensure trustworthy AI. Having said that, in the penultimate section, there are concrete examples that illustrate fictionalism about agentic AI at work, which show the explanatory power of the theory.

2. Fictionalism: theoretical background

The first relevant background feature of fictionalism involves ontic implications of our best philosophical theories. One motivation of fictionalism is to avoid ontic commitment to some kind of entity or category. In metaphysics, fictionalism has cropped up in debates about the external world, possible worlds, composite material objects, numbers, and more. Consider the existence of the external world first. The realist says that the external world exists (and is connected, causally related to the things we experience, which are conscious data); the idealist denies the existence of the external world: only data of consciousness exist. However, the idealist can employ a fictional operator to avoid committing herself to the existence of the external world as well as to avoid falling into an error theory about common sense judgements concerning the external world. She says that it is 'as if' there is an external world. In her theory, she introduces an operator 'F' that is placed before any statement of some object in the external world along with some conscious datum. Call this fictionalist idealism. According to the fiction, there is some conscious datum d that is connected to some external world object o. This is not to say that some o exists, but rather that, according to the fiction, object o exists (for critical discussion, see pp. 139-40 of [5])[1]. Consider possible worlds next. Modal realism is the view that there are possible worlds. David Lewis is its prominent defender [6], but most philosophers do not follow him. A select few of his opponents argue that modal fictionalism is true [7]. Given modal fictionalism, the actual world exists, and, according to the fiction, there are possible worlds, but the existential quantifier that quantifies over worlds inside the 'F' operator does not imply ontic commitment to possible worlds. Now consider agentic AI. The fictionalist move is to say that according to the fiction there are AI agents, but that, in reality, there are none. Any reference to or description of an AI agent is within the scope of the fictional operator. The benefit of the view is that we get to say a lot about AI agents without committing ourselves to the existence of such things. And many of the things we get to say are normative claims that govern the deployment of AI agents.

The motivation behind fictionalism differs across ontic disputes. The motivation pulls in opposite directions concerning the external world and possible worlds. In the external-world dispute, the default is realism; this is the common sense position. In contrast, in the possible-worlds dispute, the common sense position is that there exists only one real world, which is our actual world. Someone fresh to philosophical theorising will fall naturally into fictionalism about possible worlds, but resist the move to fictionalism about the external world. In contrast, in the case of agentic AI, it is not clear what the default position is. Many users of AI chatbots tend to think that the AI system or model is real, has a persona, and sometimes become emotionally attached to them[2]. Agentic AI triggers the reaction that AI agents are real because of their greater autonomy. But, on the designer/coder end and, in most technical explanations of AI technology, an AI system as a substantial thing is not taken as real; rather, it is explained that AI reacts derivatively to prompts and predicts strings of text based on complex mathematical modelling of words in a hyper-dimensional abstract space, not to mention being heavily dependent on its training data, weights, and temperature. Many users in the public must be taught that AI systems (chatbots, agents) do not think or reason like us, if what they do counts as reasoning to begin with. Nonetheless, the ordinary person's judgement should not be dismissed out of hand. Hence, whether realism or anti-realism is the default view is indeterminate.

The second relevant background feature of fictionalism concerns the different ways to draw from the notion of fiction. One way involves what Kendall Walton calls 'props', which helps to set a 'context' (for discussion, see [14]). In philosophy of art, it is noted that props set up a fiction or fictional stance. Children readily play with props in such a way that (say) the living room floor is lava. It really isn't, but according to their game it is, and given the fiction they move about in the space accordingly, jumping from one cushion to the other. This introduces pretence, which is common in acting and plays at the theatre. In 'Truth in Fiction', Lewis proposes that fictions are individuated in terms of acts of storytelling and that in the storytelling the storyteller is engaged in pretence (see pp. 265-66 of [15]).

In the case of AI, there are various technologies available on the market that naturally play into the fictionalist understanding. Character.ai has over 10 million AI characters that users can freely chat with, such as (picked at random): Aushen Ride, a motorbike enthusiast[3]. This AI character has areas of expertise, specific interests, and can draw from a wealth of information on a particular subject matter. As with other cases of fiction, under the pretence interpretation, the pretence of the AI technology is such that we appear to be told the truth about something, but really it is about or through a fictional character. This pretence need not, and often does not, involve deception on the part of the developer. The AI character is still presented as real, as uttering sentences and expressing propositions, many of which are truthful statements. Aushen Ride, for instance, when prompted, will state many truths about the best maintenance tips for motorbike riders. What is interesting about AI, and in particular agentic

[1] Outside of the fiction, fictionalist idealism includes a statement that says that conscious data exist.

[2] In some empirical studies to do with AI chatbot customer experience, users expect a certain level of humanness, because users tend to attribute humanness to the AI chatbot. See [8], [9], [10]. For discussion of anthropomorphism in marketing literature, see [11], [12], [13].

[3] https://character.ai/character/DFlK-AeE/aushen-ride-motorcycle-enthusiast

AI, is that there is an epistemic norm that what is outputted is taken to be true. In this way almost all agentic AI are fictions that are conduits to truth (for more on pretence in fictionalism, see [16]). In contrast, in many traditional forms of fiction, the content within the scope of the fictional operator is false, as in Charles Dickens telling us about Scrooge and his business exploits.

3. Agentic AI as fictional individuals

More needs to be said to flesh out the fictional approach to agentic AI. Given that this is a short paper, I shall add a few more remarks in this section to do with its application and implications. Further development of the view (both theoretical and practical) must wait for another occasion.

One factor that leads to the fictionalist understanding of generative AI, LLMs, and agentic AI is how prompt engineering is typically conceived. Prompt engineering is often regarded as an art and a science. Simple one-line prompts will result in less useful, less accurate, less interesting outputs; whereas, complex almost paragraph-length prompts will usually yield more useful, accurate, and interesting outputs. One aspect of the art of prompt engineering is building a rich context window, which requires a sophisticated use of natural language and detailed user-knowledge of the subject matter. One aspect of the science of prompt engineering is specifying the structure of a quality prompt, from designating a persona to a specific kind of task and output, etc. The structure of the prompt sets part of the context for the AI model, where the use of pretence to set a persona is explicit. Even if these trappings are not acknowledged by the user, such facts ensure the building of a fiction within which an AI model outputs content. The fiction is presupposed by design. When it comes to agentic AI, the building of a persona with guardrails, goals, and preferred methods are necessarily part of the creation of the fictional individual. These factors are even more explicit than optimal approaches to prompt engineering. So these two factors together support the case for a fictionalist understanding of generative AI models and systems, especially agentic AI[4].

I should be clear that the storytellers are the developers of agentic AI; it is real humans, coders, who create the fictional individual that is the AI agent. They are the authors of the fiction according to which there exists an AI agent that does and states various things. Perhaps, in the future, highly sophisticated AI agents (AGIs) will have the power to create AI agents, in which case those sophisticated AI agents are the storytellers. By way of further clarification, I take my fictionalism to be a semantic and pragmatic thesis, as follows. *Semantically*, AI-talk, that is, discourse about AI, is fictional. Statements about agentic AI as real, substantial things are not technically true, although within the fictional device an 'AI agent' may output true statements and execute commands that have causal efficacy. In addition, we should avoid commitment to AI agents as real things that we attribute moral properties to, such as responsibility, personhood, rights, intention, motivation, stable character traits, virtues, but *pragmatically* it is 'as if' AI agents have these moral properties. Pretending that AI agents have these moral properties serves a purpose in the development, deployment, and use of such technologies. If not, we fall into creating new kinds of harm or raising the risk of harm, for instance, users becoming emotionally attached to them.

As mentioned already, other kinds of fictionalism treat the statements in a fiction to be false. This is modelled off the standard notion of a fiction or fictional story, where fictional characters do certain things in an imagined space. Spiderman saves the day in New York City, but it is not true that he in fact saved the day. My fictionalism differs in that AI agents are recognised as attempting to state truths, and in fact perform tasks autonomously using their own decision-making and reasoning algorithms. Even if an AI agent hallucinates, the pre-defined goal is not to lie or deceive or get things wrong. The obvious examples of fictional literature to draw upon here are historical fictions, which intend to reveal truths about some time period. It is built into the context that the historical drama (say) portrays some fact. This comes out also in dramatised documentaries, which are said to be 'based on a true story'.

[4] For studies on persona prompting, see [17]; for empirical discussion on the quality of outputs using persona techniques, see [18]. For more on prompt engineering in general, see [19], [20].

4. Responsible AI according to the fiction

Now, I shall address two main questions:

1. Can AI truly be held accountable for its decisions?
2. AI as a moral agent – or merely a tool?

In answering these questions, I sketch a responsible AI framework that works well with fictionalism about agentic AI. One competing, meta-ethical theory that has been recently put forth is 'ethical behaviourism'. According to John Danaher [21], if an AI agent behaves in similar ways to entities such as human persons who have a moral status, then the AI agent should be accorded a moral status. Like traditional versions of behaviourism, behavioural outputs ground the judgement that something has some property (has a mental state, for instance). According to ethical behaviourism, then, some AI systems (AI agents) are moral agents, and not merely tools, and if so, then some AI systems (AI agents) can be held accountable for their decisions or outputs. The motivation behind this variety of behaviourism is the same motivation that leads to the Turing test. (There are other competing meta-ethical theories such as constructivism and expressivism. Given limitations on space, these other alternatives are not discussed.)

I object to ethical behaviourism. One reason is due to general problems with behaviourism already developed in philosophical literature on analytic philosophy of mind. In this literature, behaviourism is the view that mental statements and mental reports are translated into statements about behavioural outputs and dispositions. By giving a behavioural analysis of mental concepts the behaviourist claims that we need not posit a mind or internal mental states. However, as shown by D.M. Armstrong [22], Hilary Putnam [23], J.J.C. Smart [24], and others (see [25]), behaviourism does not explain all mental states (for instance, sensations, conscious experience), the analysis relies on an implausible account of dispositions, and counter-examples show that something can exhibit the requisite behaviour but intuitively we would not attribute the corresponding mental concept. To round off these problems, the upshot is that behaviourism identifies the wrong explananda; behaviour is caused by mental states, and so mental states cannot be identified with behaviour.

A second objection to ethical behaviourism has been provided from a virtue ethics standpoint [26]. Virtue ethics is centred on the moral question of what would the virtuous agent do in certain circumstances. The notion of virtue comes with the idea of the agent acting for the right reasons, motivation, feeling, and effort, with a clear recognition of why she is acting in that way. This notion of virtue is too rich to be accounted for solely in terms of behaviour. Moreover, AI systems, even agentic AI, fail to instantiate the appropriate conditions of acting or outputting for the right reasons and feeling (see p. 1550 of [26]). Therefore, an AI system cannot be virtuous in the suitable way and the behavioural analysis to reach a subject with moral status is too thin. In other words, ethical behaviourism is false.

The problem with going via the behaviour of AI systems to answer the two main questions above is that it incorrectly places moral concepts on behaviour/outputs. The attribution of moral concepts should be placed directly on the subject (that is, the AI agent), so that we say directly that the agent is responsible or performed a right or wrong action, etc. Of course, the problem with saying this directly is that it leads to reifying the AI agent. The way around this issue is to adopt fictionalism. According to the fiction, the AI agent has these moral properties and the moral properties are predicated directly. But it does not follow that the AI system or model is a moral agent. It is a moral agent only according to the fiction; however, a fiction can be useful, and there are ways to specify situations where a fictional individual is morally responsible for its outputs. So I answer that an AI agent cannot be truly held accountable, but that it can be held accountable according to the fiction, and that given the fiction we can develop a proper responsible AI framework that builds in procedural systems for when an AI agent is responsible, and of course when it is not and when it is the developer or deployer or even the user who is responsible instead.

To elaborate, one reason to accept fictionalism is its practical utility in deployment. Take for instance such AI technologies as Ana the AI Nurse by Hippocratic AI, Anna the AI emotional coach by Happify, or AI chatbots used more generally in psychiatry[5]. There are many moral dangers with deploying AI systems or models in healthcare contexts. One danger is that the AI agent is taken as a magical oracle or wizard, who knows all the facts about the science of psychology, neuroscience, psychiatry, etc. This understanding is shared by the psychiatrist and patient. The psychiatrist might defer to the judgement of the AI agent and the patient may not question its output. Both psychiatrist and patient trust the AI agent, but when something goes wrong, say, the patient is prescribed the wrong drug on the advice of the AI agent, it is not clear who is responsible. The psychiatrist did not make the decision, but the AI agent cannot be held morally responsible in the same sense as a human person. We might say that the AI agent is morally responsible, but there is no moral good in attributing blame and accountability to it. However, if we adopt fictionalism, then it is explicit in the context that the AI agent is a fictional provider of care at the outset. It is supplied in the pretence that we engage in a bit of psychiatry make-believe, and so are less inclined to think of the AI agent as some oracle who knows all and as a subject that we can attribute moral responsibility to. If we adopt fictionalism, then the use of such AI technology changes the goal to be about collaboration among AI agent, psychiatrist, and patient. This can fill the responsibility gap described above, whereby the psychiatrist and the system of care take responsibility for the fictional provider and its outputs; this is part of how the system of care is constructed by us as a practice in society.

Another example is the deployment of AI agents or AI chatbots as ‘grief bots’ (for discussion of grief bots, see [31], [32], [33]). A grief bot is an AI chatbot built upon a human person who has passed away. Surviving family members and loved ones then chat with the AI model to process their emotions. It is most clear in this use case that the context should be set up such that the grief bot is not regarded as the deceased person. One way to set up the context is to figure the deployment of the grief bot in a ritual organised by family members and loved ones (and perhaps embedded in a religious or spiritual institution). This would lead to a meaningful way to deploy AI technology to help people with grief and mourning that does not lead to a prolonged, unhealthy attachment of a loved one with the grief bot. If it is clear to the surviving person that the grief bot is a fictional individual, then this person should distinguish the fiction from the real deceased person (although this may not happen in all cases, because some people may be heavily affected by the loss, in which case those people should probably not interact with a grief bot at all). Moreover, even though it is a fiction, the person who is processing the emotion can have, if they so feel, emotions directed towards the AI agent. Frederick Kroon has recently shown how we can have emotions for fictional objects from a Brentanean perspective [34]. So fictionalism about agentic AI fits well with the use case of grief bots. As with health bots, responsibility gaps can emerge with the use of grief bots. Developing a responsible AI framework where the fiction is clear and attribution of responsibility is explicit will help fill these responsibility gaps as well.

In the previous section, I said that fiction is a conduit to truth. And it is indeed true that AI agents output facts, that is, output content (usually sentences) that express true propositions. However, as is well-known, there are errors in the outputs of generative AI models or systems. What is stated as true in the fiction may well be false. There are various explanations for this in the case of AI. First, the developer makes a mistake in the design of the AI system, either building in a poor algorithm or incorrectly curated data (biased data, for example). Second, the developer intends to deceive users who operate on shared knowledge of the AI system, that the system is trustworthy, for instance. This has a moral upshot: it puts the ethical onus back on the developer, and closes a responsibility gap where one might say that the AI agent is responsible but that it cannot be really held to account. The pretence, which is baked into the design, should not be deceptive in this way, and it should not be deceptive in portraying itself as if it is a real human being. Put differently, in any proper responsible AI framework, AI systems should be built where the context through pretence is clearly set up such that the ‘thing’ a user is interacting with is fictional and taken as such. This is especially important when users/people/human persons are in vulnerable positions such as grieving for a loved one who has passed, receiving care towards the end of one’s life, or seeking psychiatric treatment and medical relief for severe psychological problems. Semantically and fixed by convention we should agree that

[5] For discussion of AI chatbots in healthcare, see [27], [28], [29], [30].

the use of the phrases ‘AI chatbot’ and ‘AI agent’ signify a fictional individual that operates within a fiction. As a result, there is a clear path from fictionalism to ethical design practices and even AI regulation. Moreover, the fictionalist understanding of agentic AI should be explicitly stated in AI laws and regulatory standards as a common understanding that will set norms surrounding the design and deployment of agentic AI.

To sum up, fictionalism allows us to address the two main questions at the start of this section in a unified way: an AI system (agent) is not really a moral agent; in one sense it is merely a tool, but according to the fiction it is a moral agent and in virtue of that fact it can be held accountable, within the fiction. The fiction is useful for its goal in developing and sustaining responsible AI frameworks such that in certain cases an AI system (agent) will play a role in filling responsibility gaps. In other cases human persons or companies take responsibility. In the ethical domain, fictions can be taken seriously enough to align AI technologies with fundamental moral values, thereby ensuring trustworthy AI.

5. Conclusion

In this paper, I have motivated fictionalism about agentic AI and sketched a version of the theory working off extant literature on fictionalism in philosophy. According to fictionalism, agentic AI are not real entities that are attributed moral properties, but according to the fiction they exist and can be taken to have those moral properties, which serves an important purpose in developing proper responsible AI frameworks.

There are conventional practices in ethics and moral philosophy, separate from objective moral truths. It is in our power to implement and foster certain values by creating norms, principles, and practices that are adhered to. The conventional practice can be baked into the fiction. If we treat the ‘agent’ as if it is an agent, then we will take its outputs (what it does) seriously and apply rich notions of responsibility to it. This can help curtail agentic AI outputs and place moral constraints where they ought to be, whether on AI agents directly within the fiction or on human persons. Therefore, fictionalism about agentic AI is useful for creating and sustaining responsible AI frameworks that ensure trustworthy AI.

Acknowledgments

Thanks to Kirk Besmer and Tim Clancy for discussion. I also thank S. Jay Yang and funding from the Institute for Informatics and Applied Technology, Gonzaga University.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Stack Overflow, “The future is agents”: Building a platform for RAG agents, https://stackoverflow.blog/2025/05/27/the-future-is-agents-building-a-platform-for-rag-agents/, 2025. Accessed: 2025.

[2] M. Sadek, E. Kallina, T. Bohné, et al., Challenges of responsible AI in practice: scoping review and recommended actions, AI and Society 40 (2025) 199–215. doi:10.1007/s00146-024-01880-9.

[3] N. Kandhari, B. Tripathi, S. Kumar, K. Singh, N. P. Singh, Responsible AI framework for Large Language Models (LLMs), in: 2024 11th International Conference on Advances in Computing and Communications (ICACC), Kochi, India, 2024, pp. 1–6. doi:10.1109/ICACC63692.2024.10845690.

[4] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, A. Jacquet, Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering, ACM Computing Surveys 56 (2024) 1–35. doi:10.1145/3626234.

[5] D. C. Williams, Principles of Empirical Realism, Charles C. Thomas, Springfield, IL, 1966.

[6] D. Lewis, On the Plurality of Worlds, Basil Blackwell, Oxford, 1986.

[7] G. Rosen, Modal fictionalism, Mind 99 (1990) 327–354. doi:10.1093/mind/xcix.395.327.

[8] A. Rapp, A. Boldi, L. Curti, A. Perrucci, R. Simeoni, How do people ascribe humanness to chatbots? an analysis of real-world human-agent interactions and a theoretical model of humanness, International Journal of Human–Computer Interaction 40 (2023) 6027–6050. doi:10.1080/10447318.2023.2247596.

[9] L. Nicolescu, M. T. Tudorache, Human-computer interaction in customer service: The experience with AI chatbots—a systematic literature review, Electronics 11 (2022). doi:10.3390/electronics11101579.

[10] Y. Li, Z. Gan, B. Zheng, How do artificial intelligence chatbots affect customer purchase? uncovering the dual pathways of anthropomorphism on service evaluation, Information Systems Frontiers 27 (2025) 283–300. doi:10.1007/s10796-023-10438-x.

[11] E. Uysal, S. Alavi, V. Bezençon, Anthropomorphism in artificial intelligence: A review of empirical work across domains and insights for future research, in: K. Sudhir, O. Toubia (Eds.), Artificial Intelligence in Marketing, Emerald Publishing, 2023. doi:10.1108/S1548-643520230000020015.

[12] M. Blut, C. Wang, N. V. Wünderlich, et al., Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other ai, Journal of the Academy of Marketing Science 49 (2021) 632–658. doi:10.1007/s11747-020-00762-y.

[13] Y. Li, R. Hou, R. Tan, How customers respond to chatbot anthropomorphism: the mediating roles of perceived humanness and perceived persuasiveness, European Journal of Marketing 58 (2024) 2757–2790. doi:10.1108/EJM-11-2022-0827.

[14] B. Armour-Garb (Ed.), Fictionalism in Philosophy, Oxford University Press, Oxford, 2019.

[15] D. Lewis, Philosophical Papers I, Oxford University Press, Oxford, 1983.

[16] B. Armour-Garb, J. A. Woodbridge (Eds.), Pretense and Pathology: Philosophical Fictionalism and its Applications, Cambridge University Press, Cambridge, 2015.

[17] Y. Qian, Prompt engineering in education: A systematic review of approaches and educational applications, Journal of Educational Computing Research (2025). doi:10.1177/07356331251365189, forthcoming.

[18] M. Zheng, J. Pei, L. Logeswaran, M. Lee, D. Jurgens, When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15126–15154.

[19] E. Herman, Optimizing Prompt Engineering for Generative AI, Mercury Learning and Information, Boston, 2025.

[20] V. Geroimenko, The Essential Guide to Prompt Engineering: Key Principles, Techniques, Challenges, and Security Risks, Springer Cham, 2025.

[21] J. Danaher, Welcoming robots into the moral circle: A defence of ethical behaviourism, Science and Engineering Ethics 26 (2020) 2023–2049. doi:10.1007/s11948-019-00119-x.

[22] D. M. Armstrong, A Materialist Theory of the Mind, Routledge & Kegan Paul, London, 1968.

[23] H. Putnam, Brains and behavior, in: Philosophical Papers, vol. 2: Mind, Language and Reality, Cambridge University Press, Cambridge, 1975, pp. 325–341.

[24] J. J. C. Smart, Sensations and brain processes, Philosophical Review 68 (1959) 141–156.

[25] J. Kim, Mind in a Physical World, MIT Press, Cambridge, MA, 1998.

[26] M. Constantinescu, R. Crisp, Can robotic AI systems be virtuous and why does this matter?, International Journal of Social Robotics 14 (2022) 1547–1557. doi:10.1007/s12369-022-00887-w.

[27] M. Milne-Ives, C. de Cock, E. Lim, M. Shehadeh, N. de Pennington, G. Mole, E. Normando, E. Meinert, The effectiveness of artificial intelligence conversational agents in health care: Systematic review, J Med Internet Res 22 (2020). doi:10.2196/20346.

[28] S. A. Alsalamah, S. AlSalamah, H. A. Alsalamah, et al., Virtual healthcare bot (VHC-Bot): a person-centered AI chatbot for transforming patient care and healthcare workforce dynamics, Netw Model Anal Health Inform Bioinforma 14 (2025). doi:10.1007/s13721-025-00537-x.

[29] M. H. Kurniawan, H. Handiyani, T. Nuraini, R. T. S. Hariyati, S. Sutrisno, A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness, Annals of Medicine 56 (2024). doi:10.1080/07853890.2024.2302980.

[30] E. M. Boucher, N. R. Harake, H. E. Ward, S. E. Stoeckl, J. Vargas, J. Minkel, R. Zilca, Artificially intelligent chatbots in digital mental health interventions: a review, Expert Review of Medical Devices 18 (2021) 37–49. doi:10.1080/17434440.2021.2013200.

[31] J. Krueger, L. Osler, Communing with the dead online: Chatbots, grief, and continuing bonds, Journal of Consciousness Studies 29 (2022) 222–252. doi:10.53765/20512201.29.9.222.

[32] B. Jiménez-Alonso, I. Brescó de Luna, AI and grief: a prospective study on the ethical and psychological implications of deathbots, in: S. Caballé, J. Casas-Roma, J. Conesa (Eds.), Ethics in Online AI-based Systems, Academic Press, 2024, pp. 175–191. doi:10.1016/B978-0-443-18851-0.00011-1.

[33] N. F. Lindemann, The ethics of 'deathbots', Science and Engineering Ethics 28 (2022). doi:10.1007/s11948-022-00417-x.

[34] F. Kroon, Real emotions for unreal fictional objects: A brentanean perspective, Philosophia 52 (2024) 1317–1340. doi:10.1007/s11406-024-00810-9.