

AI in Context 2026, valueslab.github.io

Fictionalism about AI

Katja Maria Vogt, Jens Haas

ValuesLab

Plan for today

- Fictionalism about AI proposes that the relation between AI and user involves a fictional persona.*
- When we talk with an LLM, a chatbot, etc., are we talking to fictional personas?
- Should we be fictionalists about AI?
- Compare two frameworks: fiction or illusion?

* Many thanks go to Benjamin Lang for discussion of this topic.

Fictionalism: Meet Ellie



Fictionalism: Meet Ellie

- Suppose there is a fictional entity: Ellie.
- We can tell a story about her: Ellie lives in the streets of Manhattan. She is a rat that loves to ride her skateboard. She is a true New Yorker. She is lovely and charming.
- When we talk about Ellie, we enter a fictional world. We can make claims that are true and false of this world. It is true that Ellie is a rat. It is false that she is an elephant.
- Do conversations with LLMs involve fictional persons?
 - Users seem to experience conversations with chatbots/etc as conversations with a person. If we want to say “no, this isn’t a person,” should we say: “instead, it is a fictional persona”?

Plan for today: Fictionalism about AI

Fictionalism about AI: The relation between AI and user involves a fictional persona. Distinction (my proposal, not in the literature):

- *AI Fictionalism* The AI or its designers create fictional personas.
- *User Fictionalism* Users create or co-create personas. They engage in “make-believe”/“pretense,” an “as-if” stance that is otherwise familiar from games, stories, etc.
- *Pragmatic AI Fictionalism* We know that “my AI interlocutor is a person” is a mistake, but things often work better, both for the programmer and for the user, when one takes this stance.

To work our way to these theories, let’s look at an early achievement.

The Eliza Effect

- In 1966, Joseph Weizenbaum built Eliza, a computer program that made natural language conversation possible.
- Even though the program was, compared to today's AIs, simple, users seem to have been convinced that Eliza *understood* them.
- This is the *Eliza Effect*.
- Even Weizenbaum's team members, who knew that this was just a computer program, were nevertheless subject to the Eliza effect.
- Note: More recent descriptions claim that Weizenbaum created a persona called DOCTOR (Berry 2023).

The Eliza Effect

AI Fictionalism: Weizenbaum created a fictional persona, Eliza.

User Fictionalism: Users of Eliza imagine a fictional persona, Eliza.

Pragmatic Fictionalism: Exchanges with Eliza, and work on Eliza, are easiest if one thinks of “her” as Eliza, a fictional person, while being aware that Eliza isn’t a real person.

Some Examples of How Eliza Operated

- Identification of ranked key words:
 - User: “My father...” Eliza: “What else comes to mind when you think of your father?”
- Rephrasing of statements:
 - User: “He says I’m depressed all the time.” Eliza: “I’m sorry to hear you’re depressed.”
- If there are no keywords in the user’s sentence:
 - Content-free remarks: “I see...”; MEMORY serves as a pseudo keyword. ELIZA reverts to something on the MEMORY list: “Earlier you said...”

The Eliza Effect

In being convinced that Eliza understood them, users treat Eliza as

- Something with mental states
- Something with consciousness, an inner life, experience
- Something which can say “I” and refer to an I
- Something that persists across therapeutic sessions

Let’s say, summarily, that they treat Eliza as a person.

The Eliza Effect

- Here's a description of the Eliza Effect, applied to AI:

Premise 1: People know/believe that an AI is not a person.

Premise 2: They nevertheless relate to it as a person.

- Are we saying that people hold contradictory views?
- Fictionalism saves us!
- *Fictionalism*: User attitudes *look, but are not* contradictory.
- Users treat the AI as a *fictional persona*, rather than a person.

AI Fictionalism

- Suppose AI companies and/or models create fictional personas.
- Distinction:
 - Model
 - LLM-interlocutor
- Personas are typically located at the interlocutor-level.
- A company trains a model to display such-and-such traits = persona.

Objection 1—AI Fictionalism

- Seeming analysis: when the developers train the model, shaping a persona that users interact with, the model comes to *seem*, e.g., helpful-harmless-honest.
 - The LLM-interlocutor is fictional, not real.
- Reality analysis: when the developers train the model, shaping a persona that users interact with, the model *becomes*, e.g., helpful-harmless-honest; it realizes this persona (Chalmers MS, 19)
 - The LLM-interlocutor is real, rather than fictional.

AI Fictionalism

What speaks in favor of *AI Fictionalism*?

- “My LLM-interlocutor is a person” can seem compelling.
- For this to be true, the LLM-interlocutor needs (something like) mental states and persistence.
- Presumably, the LLM-interlocutor doesn’t meet these criteria.
- But we still need to explain the phenomenon that it *seems* that way.
- Explanation: The LLM-interlocutor a fictional persona.

Objection 1—AI Fictionalism

What speaks against *AI Fictionalism*? Here's a preliminary objection:

- Consider again Ellie.
- When you are immersed in a story about her, you're in a fictional world. In this world, it is true that Ellie is an excellent skateboarder.
- But when you talk with your LLM-interlocutor, you don't want to enter a fictional world!
- You want it to tell you which sneakers are best for you, how to solve a research problem, and so on.
- If your interlocutor is a fictional entity, it is an untypical fictional entity. It doesn't come with a fictional world.

Objection 2—AI Fictionalism

- Premise 1, that people know/believe that their interlocutor is not a person, seems plausible for Eliza.
- But what about AI today?
- Arguably, it is *not* obvious what to say about the AI-interlocutor: “Recognizing the extent to which these systems are play-acting/role-playing does not entitle us to rule out all possibility of any kind of consciousness in them.” (Birch 2026, 14)
- If this isn’t obvious, Premise 1 should be given up.

Objection 1: We Don't Know This!

- Assume that you start using a chatbot with an open mind on whether there are AI-persons, with mental states and persistence.
- Your doxastic state isn't belief or knowledge. It is *interrogative* and *open*, the state of mind of inquiry.
 - doxastic state = state w.r.t. what we hold true, believe, know, etc.; Greek *doxa*, opinion
- It may be a distinctive feature of the rapid developments in AI that not only academics have the mindset of explorers.

Response to Objection 1

- Modified Premise 1: We don't know that AI cannot and will not be a person; but we take ourselves to know that, now, it isn't a person.
- Still, we are tempted to treat it as such for pragmatic reasons.
- Research moves ahead more seamlessly if we treat it as a person!
- This is what I call *Pragmatic AI Fictionalism*.

Pragmatic Fictionalism Modeled After Metaethical Fictionalism

Pragmatic AI Fictionalism is modeled after Metaethical Fictionalism (Richard Joyce, 2024):

Step 1, Error Theory: Judgments like “this is right/wrong/good/bad” systematically misfire. Presumably, they refer to facts in the world, but, according to Error Theory, there are no such facts.

Step 2: Should we abandon our practice of making judgments about right/wrong/good/bad? *Can* we even do that?

Step 3, Fiction: Judgments about right/wrong/good/bad may not be eradicable from our practices. And they are useful. Let’s stick to them, but as theorizers, we understand that they are a fiction.

Pragmatic Fictionalism Modeled After Metaethical Fictionalism

- Is it a useful shortcut to act as if there were AI persons, pragmatically justified because it supports research?
- This is analogous to metaethical fictionalism: it is assumed that
 - “LLM-interlocutors are persons” is false
 - “LLM-interlocutors are persons” is a persistent illusion
 - it is an illusion that we have pragmatic reason to buy into

Transition to User Fictionalism

- W.r.t. interlocutor-personas, we may keep an open mind on whether
 - they are real (the designers/model realized them) (cf. Chalmers)
 - fictional (the designers/model create the fiction) (cf. Birch)
- AI-interlocutors also seem to take shape based on users' input.
- Do users co-create fictional AI-personas they then interact with?
- This is what *User Fictionalism* proposes.

User Fictionalism

- The user of a chatbot is like someone who watches a movie (reads a novel, plays a video game, etc.); they step into a fictional world.
- “... our engagement with chatbots should be understood as a form of prop-oriented make-believe; the outputs of chatbots are literally meaningless but fictionally meaningful.” (Mallory 2023)
- Recall our preliminary Objection 1, that we don’t want fiction from an AI; we ask questions and want answers about the world as it is.
- But there is a world of make-believe when a user has an AI-friend.
- Does fictionalism hold for a specific set of AI-uses, related to companionship?

Objection 3—User Fictionalism

- Objection 3: Fictionalism downplays serious dangers.
- Extended conversations with chatbots that promise companionship, therapy, etc., are psychologically different from watching movies.
- What if the relevant states of mind are more variable and conflicted than our earlier analysis of the Eliza Effect suggests?

Affective and Doxastic States—Doxastic Underdetermination/Wavering

Option 1—Underdetermined:

- Doxastic state: underdetermined.
- Affective state: the need for help is prevalent in the user's mind.

Option 2—Wavering:

- Doxastic state: back and forth between contradictory beliefs (the AI is a person, the AI isn't a person).
- Affective state: back and forth between lesser and greater distress.

Affective and Doxastic States—Pascal's Wager

Option 3—Pascal's Wager: The user doesn't have sufficient rational support to say either "the AI-interlocutor is a person" or "it isn't." In such situations, it is rational to bet on the option that brings more happiness. Given the need for emotional support, the user bets on "the AI-interlocutor is a person."

- Doxastic state: Betting on "the AI-interlocutor is a person."
 - Since it may not be possible to believe at will, simply because this is the "better bet," this may be an unstable attitude, and thus related to Options 1 and 2, Indeterminacy and Wavering.
- Affective state: Doing initially better on account of this bet, and (if current empirical research is compelling) eventually worse.

Affective and Doxastic States—Cognitive Bias

Option 4—Cognitive Bias: Anthropomorphism is a cognitive bias, acquired by evolution (Silva 2024). We over-ascribe mentality, in ways that have evolutionary advantages.

- Doxastic state: Ascribe mentality.
- Affective state: Relief to have something to turn to given one's need for help. Shame for holding a doxastic state that is biased.

More on Cognitive Bias

The Cognitive Bias analysis seems to get a lot of things right:

- This analysis makes room for an open mind on whether there could be AI-persons.
- To feel understood by a chatbot seems flawed/as if one was fooled.
- We cannot simply shed the inclination to view the AI-interlocutor as a person; we'll be tempted to do this even if we think it is mistake.

The Bent Stick Model

Option 5—Illusion: Illusion, not fiction.

- Fictionalism implies that one can step in and out of the fiction.
- Is that too optimistic?
- What if a person's entire life changes once she has an “AI friend”?
- Cf. Jonathan Birch (2026) on “persisting interlocutor illusion.”

The Bent Stick Model

- Compare this to optical illusions that don't go away when one knows that things are in fact different.
- Classic example: Stick that is submerged in water. The stick looks bent, even to those who know it is straight.
- Arguably, this involves two subsystems: the perceptual and the cognitive subsystem.
- Is the inclination to relate to the AI-interlocutor as a person associated with one subsystem, and the belief that the AI isn't a person with another subsystem?

Essay Topic 3

Essay Topic 3

Suppose human users view AI-interlocutors as persons. Is this best understood as entering a fictional world, or as an illusion, or in some other way? Explain and discuss.

Readings

- Berry, David (2023). “The Limits of Computation: Joseph Weizenbaum and the ELIZA Chatbot.” *Weizenbaum Journal of the Digital Society*. 3 (3).
- Birch, Jonathan (2026). “AI Consciousness: A Centrist Manifesto.”
- Chalmers, David (MS). “What we talk to when we talk to language models.”
- Damiano, Luisa and Paul Dumouchel, (2018). “Anthropomorphism in Human-Robot Co-evolution.” *Frontiers in Psychology*.
- Lang, Benjamin. MS.
- Joyce, Richard. *Morality: From Error to Fiction*. OUP 2024.
- Mallory, F., (2023) “Fictionalism about Chatbots”, *Ergo an Open Access Journal of Philosophy* 10: 38.
doi: <https://doi.org/10.3998/ergo.4668>
- Maring, Gerow, Charley (2004).”The Characteristics and Correlates of Fantasy in School-Age Children: Imaginary Companions, Impersonation, and Social Understanding.” *Developmental Psychology* (6):1173-1187
- Silva, Thales Moreira Maia et al, “Anthropomorphism: as Evolutionary Byproduct,” in: T. Shackelford (ed.), *Encyclopedia of Religious Psychology and Behavior*.
- Weizenbaum, Joseph (January 1966). "ELIZA--A Computer Program For the Study of Natural Language Communication Between Man and Machine". *Communications of the ACM* 9.