Mentalistic stances towards AI systems: beyond the Intentional stance

Silvia Larghi[1][0009-0006-2749-1411] and Edoardo Datteri[1][0000-0003-0323-2985]

[1] RobotiCSS Lab, Laboratory of Robotics for the Cognitive and Social Sciences, Department of Human Sciences for Education, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, Milan, Italy

Abstract. Under what circumstances do we attribute a mind to AI systems? And, in this case, how do we think their mind works? Answering these questions is crucial to inform the design of safe and trustable AI, to inform research on the ethical, social and legal issues raised by the increasing presence of AI systems in everyday life and to investigate how they can be used as tools to study human and social cognition. This work proposes a philosophical reflection on the possible structure of people's mental models of AI systems. We distinguish between two possible styles of modelling that people may adopt in everyday contexts. Both involve the attribution of mental states and cognitive abilities to the AI system, even though they differ from one another in some relevant aspects. One modelling style is akin to folk psychology and relies on our commonsensical concept of mental representation. The other, which we will refer to as folk-cognitivist, is more akin to the account of the structure of the mind that characterises classical cognitive science. These modelling styles correspond to different classes of mentalistic stances that people may adopt when they interact with AI systems in ordinary contexts.

Keywords: philosophy of Artificial Intelligence, philosophy of Cognitive Science, human-AI interaction, mental state attribution.

1 Introduction

It has been recently claimed that Generative Artificial Intelligence (AI) systems, such as Open AI's ChatGPT, can be *hypnotized* (Sharma, 2023). Whether this is the case or not, it is interesting to note that a psychological term denoting a cognitive alteration is used to characterise aspects of the functioning of AI systems of a particular sort. Similarly, situations where the content generated by Large Language Models (LLMs) results "nonsensical or unfaithful to the provided source content" are called *hallucinations* (Ji et al. 2023). Claims of this sort should come as no surprise. Indeed, today's AI systems often display characteristics, like the striking fluidity of natural language interaction and the consistency of the generated texts, which may be expected to induce people to occasionally attribute mental states and other cognitive abilities to them. Consistently, in a growing number of research studies, paradigms, methods and constructs used to study human psychology (including the theory of mind) are applied

to the modelling of LLMs *psychology* (Brunet-Gouet et al., 2023; Kosinsky, 2023; Loconte et al., 2023). A case in point is Dietz and colleagues (2023), who studied adults' and children's understanding of the *mind* of conversational agents such as smart speakers.

These considerations prompt questions that have an empirical and a philosophical side. Under what circumstances do we attribute a mind to AI systems? And, in this case, how do we think their mind works? Answering these questions is important to inform the design of safe and trustable AI (Ziemke, 2020; De Graaf & Malle, 2017), to inform research on the ethical, social and legal issues raised by the increasing presence of AI systems in everyday life - from social robots to conversational agents and other sorts of virtual or embodied agents (Carrillo, 2020; Sullivan et al., 2022; Jaeger and Levin, 2016), and to investigate how they can be used as tools to study human and social cognition (Wykowska, 2021; 2020).

This work proposes a philosophical reflection on the possible structure of people's mental models of AI systems. More specifically, we distinguish here between two possible styles of modelling that people may adopt in everyday contexts. Both involve the attribution of mental states and cognitive abilities to the AI system, even though they differ from one another in some relevant aspects. One modelling style is akin to folk psychology and relies on our commonsensical concept of mental representation. The other, which we will refer to as *folk-cognitivist*, is more akin to the account of the structure of the mind that characterises classical cognitive science. The main claim made in this paper is that these modelling styles correspond to different classes of mentalistic stances that people may adopt when they interact with AI systems in ordinary contexts. The analysis presented here is part of a wider research project, whose future step will involve the refinement of the distinctions made here, and the development of theoretical frameworks through which the "on-line" verbal and non-verbal human-AI interactions can be analysed, for the purpose of gaining a deeper understanding of people's mental models of AI systems.

The structure of the paper is as follows. The two modelling styles introduced here will be presented in detail in section 2, based on a review of recent studies on the attribution of mental states to artificial systems. Section 3 will provide some examples. Section 4 will provide a summary and concluding remarks.

2 Mentalistic stances towards AI systems

2.1 Mental state attribution to AI systems

The attribution of mental states to artificial agents has been widely studied in human-robot interaction and human-computer interaction. Thellman and colleagues (2022) review a rich list of studies on mental states attribution to robots carried out from several diverse perspectives, including psychology, neuroscience, computer science, and philosophy. This literature starts from the presupposition that people may adopt either a mentalistic or a non-mentalistic stance towards artificial agents while explaining and predicting their behaviour (De Graaf & Malle, 2017). While mentalistic explanations of behaviour refer to the mind and mental capacities of the agent (De Graaf

& Malle, 2019), non-mentalistic explanations of behaviour do not refer, either implicitly or explicitly, to the systems' mind: they are typically (but not necessarily) based on the theoretical vocabulary of physics and/or electronics (e.g., the robot is stuck because the battery is low).

Notably, the studies published so far on the *mentalistic* explanation of artificial agents' behaviour heavily rely on Dennett's conceptual framework (1971, 1987). Famously, according to Dennett, there are several possible stances one can adopt to explain and predict the behaviour of a system: the physical stance (where one explains the systems' behaviour with reference to its physical states), the design stance (which refers to the functional design of the system), and the so-called intentional stance. The latter stance consists in explaining the system's behaviour by assuming that it is rational and ascribing beliefs, desires, intentions and other intentional states to it. Contemporary literature on mental states attribution to artificial agents chiefly relies on this framework (Thellman et al., 2017; Marchesi et al., 2021; Thellman & Ziemke, 2019). Perez Osorio and Wykowska (2020) provide a review of the many studies trying to assess whether and in what conditions people adopt the intentional stance towards robotic systems based on empirical tools such as the questionnaire proposed by Marchesi and colleagues (2019). Is the intentional stance the *only* kind of *mentalistic* stance that people may adopt towards artificial systems in ordinary interactions? We believe it is not, as explained in the rest of this paper.

2.2 Styles of mentalistic modelling and mentalistic stances towards AI systems

To pave the way for the ensuing discussion, it is important to clarify what is meant here with "stance". We start from the assumption that, when people interact with other agents, they form mental models of them and use these models to explain and predict the agent's behaviour. In our perspective, the formulation of a mental model of an AI system (and its explanatory and predictive use) amounts to "taking a stance" towards them.

The notion of mental model has been explored and discussed by several scholars (most notably, Johnson-Laird, 1983). Here we will construe this notion along the lines of Achinstein's analysis of theoretical models in physics (Achinstein, 1965). According to Achinstein, theoretical models possess some characteristics that chiefly include the following:

1) "A theoretical model consists of a set of assumptions about some object or system".

2) "A theoretical model describes a type of object or system by attributing to it what might be called an inner structure, composition or mechanism, reference to which will explain various properties exhibited by that object or system."

3) "A theoretical model is treated as an approximation useful for certain purposes", implying that there may be alternative models in use.

We construe the notion of a "mental model" of an AI system along these lines. More specifically, in the framework proposed here, mental models of AI systems can be conceived as sets of beliefs, possessed by the modeller, whose contents express a number of assumptions about the AI system. These assumptions state that the AI sys-

tem has a particular structure, composition or mechanism. Reference to this structure can be used to explain and predict the behaviour of the AI system. Moreover, mental models capture some aspects and not others of the modelled system. Whereas Achinstein uses the term "approximation", we prefer to use the terms "abstraction" and "idealization" that are often used in the philosophical literature on models (see Frigg and Nguyen, 2017) to refer to the omission of certain aspects and the introduction of falsities, respectively, in the modelling of the target system.

Our conception of mental model refers to the notion of "belief". This notion is used here with the classical meaning of a propositional attitude, defined by having a certain attitude (believing) towards a content expressed by a proposition (for a concise discussion of this classical interpretation, see Crane, 2016). If an agent A holds the belief that another agent B is hungry, then A has an attitude (believing) towards a content expressed by the proposition "B is hungry". A may have different attitudes towards the content expressed by the proposition "B is hungry". For example, A might *desire* that B is hungry. Beliefs can play psychological roles (Fodor, 1975) influencing the actions of the believer. For example, agent A's belief that agent B is hungry may induce A to feed B. We are thus claiming that, while interacting with an AI system, the user may form mental models of the system. These mental models can be understood as sets of beliefs held by the user whose content somehow refers to the modelled system. Some of these beliefs will attribute an inner structure, composition or mechanism to the system. For the sake of generality, let us assume that the content of these beliefs can be represented in a canonical form as "the AI system S has characteristic X".

Now, one may surely *believe* that the AI system has certain *beliefs*. More specifically, the user's mental model may include beliefs stating that the AI system is itself characterised by the possession of certain beliefs, desires, intentions, or propositional attitudes of various sorts. We will call this style of mental modelling 'folk-psychological'. Classically, folk psychology, or common-sense psychology consists in the attribution of beliefs, desires and other propositional attitudes to other agents, plus law-like generalisations, such as: If someone wants X and holds the belief that the best way to get X is by doing Y then, *ceteris paribus*, the person will do Y (for general discussions on folk psychology, see Ramsey, 2007; Jackson & Pettit, 1990; Horgan & Woodward, 1985; Stich & Ravenscroft, 2011; Stich 1983). Note that Dennett's intentional stance can be readily accommodated within this framework. According to Dennett, to adopt the intentional stance towards a system consists in "ascribing to the system the possession of certain information and by supposing it to be directed by certain goals, and then by working out the most reasonable or appropriate action on the basis of these ascriptions and suppositions" (Dennett, 1971), where the discussion following in Dennett's seminal article makes it clear that the system's possession of certain information can be equated to the possession of certain beliefs and desires.

A long-standing debate in the philosophy of mind and science concerns the distinction between folk psychology and cognitive science (a classical discussion being made by Stich, 1983). Whether a deep distinction exists between the two is out of the scope of this paper. We rather claim that the two approaches to the modelling of the mind differ at least *superficially*. Where folk psychology models the mind in terms of propositional attitudes and law-like generalisations among them, *prima facie*, cogni-

tive theories adopt a different theoretical vocabulary, in which the notions of representation and information-processing modules play a central role. In cognitive theories, the mind is modelled in terms of cognitive modules which perform information-processing functions, typically characterised by I/O relationships, where the inputs and the outputs are representations of external or internal states (Bechtel, 2008; Cummins, 1983; Fodor, 1983; Pylyshyn, 1984). A paradigmatic case in point is Marr's well-known theory on visual perception (Marr, 1982).

Our contention here is that not only cognitive scientists, but also laypeople, may occasionally form mental models of AI systems that ascribe to them not beliefs and rationality, but rather a set of information-processing modules and representational structures. In this case, we say that they form a *folk-cognitivist stance* towards the AI system.

3 Some notional examples

To illustrate, consider the following scenario. John asks a smart speaker to play Rihanna's "Umbrella", but the system, in response, plays Bob Dylan's "Visions of Johanna". How will the user explain this behaviour? John might adopt what we have called a folk-psychological modelling style. They might hold (in their mental model) the belief that the system believed that the title of Bob Dylan's song was "Umbrella"; or they may traffic at the second-order level and believe that the smart speaker believed (incorrectly) that the user wanted to listen to "Visions of Johanna". A bystander might form a mental model of the smart speaker and believe that the system believes that Rihanna's songs are not that good and that John must change their musical taste. And so on. In all these notional examples, John and the bystander hold a mental model of the smart speaker which is couched in folk-psychological terms. Or, in Dennett's terms, they are taking an intentional stance towards the system.

On the other hand, the user might adopt what we have called a folk-cognitive modelling style. In this case, the user would believe, for example, that the smart speaker has a certain mechanism for processing the vocal commands of the user, and that something went wrong after John's request for Rihanna's song: the smart speaker did not correctly decode John's audio input. More verbosely, the user's beliefs about the system might have the following contents. The smart speaker has a sensor (a microphone) which stores the raw input audio data (the user's vocal request) in a part of the memory. Then, there is some information-processing module in the system that transforms the input audio data into a textual representation of the vocal command. The textual representation of the command is then sent to other functional modules which perform a web search of the corresponding song; and so on and so forth. For some reason, the raw input data (flowing from John's utterance "Play Rihanna's 'Umbrella'!") were transformed into the text "Visions of Johanna by Bob Dylan". Or, the decoded text was actually "Umbrella by Rihanna", but the web search returned Bob Dylan's song. Let us examine some characteristics of this model.

First, there is no reason to exclude that John's mental model of the system could take this form. Mental models, in our framework, are sets of beliefs about the system.

In this case, John would believe that the smart speaker has a number of characteristics - an inner structure of mechanism. Second, at least prima facie, this model does not attribute beliefs, desires, intentions to the smart speaker, as folk-psychological explanations would do. Believing that the system is able to process information is different from believing that the systems has beliefs, desires, and intention, to the same extent and in the same way cognitive science theories are different from folk-psychological theories. This is not be a folk-psychological mental model of the functioning of the smart speaker. It does not result from taking the intentional stance towards the system.

Third, this would be a mentalistic model for the same reason that cognitive theories model (people's) minds. It refers to information-processing modules that process representations (see Figure 1). With the possible exception of the microphone and the speaker, no reference is made to the physical structures that eventually implement (whatever this means) these information-processing modules. There is substantial disagreement as to what constitutes cognition (see for example Adam and Aizawa, 2001), yet if cognition is computation over representational states, then arguably the smart speaker, according to this model, possesses at least one of the important requirements for being a cognitive system. The folk-cognitivist model discussed here models the mind of the smart speaker in the same sense information-processing theories in cognitive science are typically told to model aspects of people's mind.

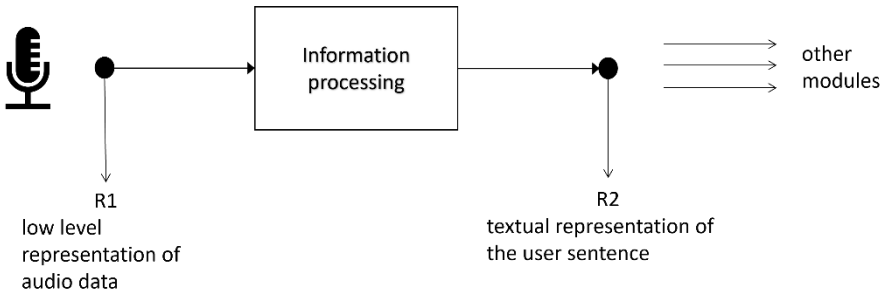

Fig. 1. A possible internal structure attributed to the smart speaker according to the folk-cognitivist modelling style

4 Taking Stocks

In this paper we have distinguished between two styles people may adopt to model the functioning of AI systems, which correspond to two possible classes of mentalistic stance that people can adopt in explaining and predicting the AI systems behavior. One modelling style is folk-psychological and involves the attribution of rationality and beliefs, desires and other propositional attitudes to the AI system. The other modelling style has been dubbed folk-cognitivist: it is based on the ascription to the AI system of information-processing and representational abilities. The main takeaway from this work is that people, at least in principle, in their everyday interaction with AI systems, may form mental models of (i.e., take mentalistic stances towards) them

that differ from the intentional stance and more closely resemble cognitivist styles of explanation.

This claim gives rise to a number of philosophical and empirical questions that will be addressed in future research. On the philosophical side, one may object that the folk-cognitivist stance sketched here is nothing more than Dennett's design stance. Even though, admittedly, more work is needed to fully reveal the differences between the design and the folk-cognitivist stance, some prima facie differences exist between the two. In Dennett's framework, neither the design nor the physical stance - but only the intentional stance - ascribe intentionality to the system. The folk-cognitivist stance, instead, exactly consists in believing that the system has a number of intentional entities or states, viz. internal representations that are "about" external or internal states of affair. In other words, whereas in Dennett's framework intentionality is the mark of the intentional stance and is lacking in the design stance, both the folk-psychological and the folk-cognitivist modelling style assume the intentionality of the system. Moreover, whereas the notion of "function" plays a central role in the design stance as characterised by Dennett (1971, p.88), folk-cognitivist theories may also be couched in non-teleological terms, e.g., in terms of input-output regularities or the algorithms underlying them.

More empirical problems arising from the discussion made here are: how plausible is that laypeople - in particular, people who are not expert in cognitive science - adopt the folk-cognitivist style? And what methods could be deployed to assess what style people adopt in different situations? As far as the first question is concerned, we note that concepts belonging to computer and cognitive science, like 'information', 'data', 'memory', 'representations' and similar, have become part of our everyday pre-theoretical vocabulary. It would come as no surprise if they were already inflating laypeople's models of AI system. But this is only a suggestion, that must be evaluated empirically - which leads us to the second question. What methods should be used to study people's mental models of AI systems? Arguably, people's verbal utterances cannot be taken to always express their beliefs literally. John's yelling "You did not understand my request!" does not imply that John literally believes that the system did not understand their request - it could be a metaphorical saying. However, one may devise experimental situations in which folk-psychological and folk-cognitivist models lead to different observable predictions on the behaviour of the user. Notably, these predictions might be about the verbal as well as the non-verbal behaviour of the user during prolonged interaction with the system. These questions are shaping an ongoing research project that will hopefully lead to a further step towards understanding how people understand AI systems.

References

Achinstein, P. (1965). Theoretical Models. *The British Journal for the Philosophy of Science*, 16(62), 102–120.

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical psychology*, *14*(1), 43-64.

Bechtel, W. (2008). Explanation: Mechanism, modularity, and situated cognition. *The Cambridge handbook of situated cognition*, 155-170.

Brunet-Gouet, E., Vidal, N., & Roux, P. (2023). Do conversational agents have a theory of mind? A single case study of ChatGPT with the Hinting, False Beliefs and False Photographs, and StrangeStories paradigms. *Zenodo*

Carrillo, M. R. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, *44*(6), 101937.

Crane, T. (2016). *The Mechanical Mind. A Philosophical Introduction to Minds, Machines and Mental Representation* (3rd ed.). Routledge

Cummins, R., (1983) *The nature of psychological explanation*, MIT Press, Cambridge, Mass.

De Graaf, M. M., & Malle, B. F. (2017, October). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

De Graaf, M. M., & Malle, B. F. (2019, March). People's explanations of robot behavior subtly reveal mental state inferences. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 239-248). IEEE.

Dennett, D. C. (1971). Intentional systems. *The journal of philosophy*, *68*(4), 87-106.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dietz, G., Outa, J., Lowe, L., Landay, J. A, & Gweon, H. (2023). Theory of AI Mind: How adults and children reason about the “mental states” of conversational AI. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45.

Fodor, J. A. (1975). *The language of thought*, Harvard University Press, Cambridge (Mass.).

Fodor, J. A. (1983). *The modularity of mind*. The MIT press.

Frigg, R., & Nguyen, J. (2017). Models and representation. *Springer handbook of model-based science*, 49-102.

Jackson, F., & Pettit, P. (1990) In defence of folk psychology, *Philosophical Studies*, *59*(1), 31–54.

Horgan, T., & Woodward, J. (1985) Folk Psychology is Here to Stay, *The Philosophical Review*, *94*(2), 197.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38.

Jaeger, C. B., & Levin, D. (2016). If Asimo thinks, does Roomba feel? The legal implications of attributing agency to technology.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge University Press.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Loconte, R., Orrù, G., Tribastone, M., Pietrini, P., & Sartori, G. (2023). Challenging ChatGPT 'Intelligence' with Human Tools: A Neuropsychological Investigation on Prefrontal Functioning of a Large Language Model. *Intelligence*.

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots?. *Frontiers in psychology*, *10*, 450.

Marchesi, S., Spatola, N., Perez-Osorio, J., & Wykowska, A. (2021, March). Human vs humanoid. A behavioral investigation of the individual tendency to adopt the intentional stance. In Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (pp. 332-340).

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*, W. H. Freeman, San Francisco.

Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, *33*(3), 369-395.

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. MIT Press.

Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.

Sharma, S. (2023), LLMs like GPT and Bard can be manipulated and hypnotized, https://interestingengineering.com/science/llms-like-gpt-and-bard-can-be-manipulated-and-hypnotized, last visited on 15 September 2023

Stich, S. P. (1983) *From Folk psychology to cognitive science: The case against belief*, The MIT Press.

Stich, S., & Ravenscroft, I. (2011) What is Folk Psychology?, in: *Collected Papers*, Volume 1, Oxford University Press, pp. 214–234.

Sullivan, Y. W., & Fosso Wamba, S. (2022). Moral judgments in the age of artificial intelligence. *Journal of Business Ethics*, *178*(4), 917-943.

Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in psychology*, *8*, 1962.

Thellman, S., & Ziemke, T. (2019). The intentional stance toward robots: Conceptual and methodological considerations. In *The 41st Annual Conference of the Cognitive Science Society, July 24-26, Montreal, Canada* (pp. 1097-1103). Cognitive Science Society, Inc..

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

Wykowska, A. (2020). Social robots to test flexibility of human social cognition. *International Journal of Social Robotics*, *12*(6), 1203-1211.

Wykowska, A. (2021). Robots as mirrors of the human mind. *Current Directions in Psychological Science*, *30*(1), 34-40.

Ziemke, T. (2020). Understanding robots. *Science Robotics*, *5*(46), eabe2987.