This is the author's eprint of a chapter in In A. Strasser (ed.), Anna's AI Anthology. How to live with smart machines? (pp. 55–78). Xenomoi, 2024. It may differ in minor ways from the print version, which is the definitive one. Version of 1/10/24.

What are Large Language Models Doing?

Keith Frankish

***Abstract**: Are they just spewing out text mechanically, like a parrot or a printer, or are they genuinely conversing with us, answering our questions, complying with our instructions, and so on? Do large language models (LLMs) perform intentional actions? Do they possess beliefs and desires that motivate the replies they produce? Treating LLMs as intentional systems gives us considerable predictive power. Yet, at the same time, it is implausible to think that LLMs possess communicative desires and perform speech acts. So, what are LLMs really doing?*

> The meaning doesn't matter if it's only idle chatter of a transcendental kind. – *Patience, W. S. Gilbert*

1. Is there a philosopher in the house?

I imagine that if you graduate from medical school, then you secretly hope that one day your skills will be called upon in some dramatic way. On a plane or at the theatre, the call will go out for a doctor, and you will shout, "I am a doctor!". And you will step in and save the day. I doubt if many philosophers harbour such fantasies, but this might be the moment when society makes a similar dramatic call upon their services. For recent developments in artificial intelligence raise questions that are distinctively philosophical in nature.

I am thinking, of course, of large language models (LLMs), such as OpenAI's GPT-3 and GPT-4. These are massive neural networks that have been trained to perform text-completion tasks on vast quantities of human-produced text. Linked to a chatbot application such as OpenAI's ChatGPT, these models can produce linguistic responses to queries and instructions that appear to display a remarkable level of knowledge and intelligence. They seem to hold conversations, offer advice, write essays, pass exams, and much more. Note that I say that they *seem to* do these things; whether they are really doing them will be the topic of this chapter. (For convenience, I shall use the term "LLM" to mean a large language model integrated with a chatbot.)[1]

The existence of such systems poses many questions. Some are social ones – ethical, economic, legal, and political. How are we going to live with this new technology? What impact will it have on our economy and society? Will LLMs take over jobs and social roles formerly occupied by humans? And if so, will

[1] A good introduction to LLMs and how they work can be found on Stephen Wolfram's website (Wolfram, 2023).

they perform the tasks reliably, or will they spread misinformation, polluting the sources of knowledge on which we rely? Will their apparent intelligence induce us to place too much trust in them, opening ourselves to error and exploitation? Should we regulate their development and use, and if so, how? These are urgent questions, and they are questions for everyone, but philosophers, with their expertise in ethics and political theory, should have a lot to contribute to the debate.

There are also challenges of a more theoretical nature, involving metaphysical and conceptual questions. What kinds of things are these smart machines we have created? Do they have minds with beliefs, desires, and intentions? Do they understand what they are saying? Might they be, or become, conscious? Philosophers of mind should be able to help with these questions, at least by clarifying the issues, outlining the theoretical options, and setting out the arguments for and against the various positions.

2. Intentional action

The main topic for this chapter is a question of the second kind. It's the question of what LLMs are *doing*. Of course, in one sense we already know what they are doing: they are generating sentences in a human language (linguistic outputs) in response to prompts we type in. But is that all they are doing? Are they just spewing out text mechanically, like a parrot or a printer, or are they genuinely *conversing* with us, answering our questions, complying with our instructions, and so on?

What exactly is the difference, anyway? After all, our own utterances are produced by neural mechanisms of some sort. We might say that genuine conversation requires an *understanding* of what is said. That's true enough, but understanding isn't easy to define or quantify. (Do I understand the sentence, "In the forest, there were elm, ash, and beech trees"? Sort of, though I couldn't identify those trees or tell you what distinguishes them.) Instead, I shall focus on a related but more tractable question. When we converse, we do so *for a reason*. We ask questions because we *want* to know something and *believe* that our hearer may know the answer. And when we answer others' questions, we do so because we *want* to help and *believe* that the words we utter express what they want to know. Philosophers call actions like these, which are done for a reason, *intentional actions*.[2] ("Intentional" here is a philosophers' technical term; it means *about* or *directed to* something – a goal in this case.) Now, if humans were to produce linguistic outputs similar to those of LLMs, we would

[2] For a detailed examination of different notions of action and their application to LLMs, see Joshua Rust's "Minimal agency in living and artificial systems" (this volume).

take them to be performing intentional actions. We would assume that they want to give cooperative responses and believe that their words express such responses. Does the same go for LLMs? Are they also performing intentional actions motivated by beliefs and desires? And if so, exactly what intentional actions are they performing, and what are the beliefs and desires that motivate them?[3]

These are the questions I shall be addressing. I shall adopt a policy of being as concessive as possible to LLMs, adopting the theoretical approach most likely to yield the verdict that they *do* possess mental states and perform intentional actions. To anticipate my conclusion, I shall argue that, even on this generous interpretation, LLMs possess only a limited range of mental states and perform only one type of intentional action. This conclusion will have implications for questions of the first type I mentioned – questions about the risks LLMs pose – and I shall discuss these briefly at the end of the chapter.

3. Propositional attitudes

How do we tell if a system has beliefs and desires and performs intentional actions? (I shall use the term "system" to include both organisms and artificial devices.) We shall need to put some background in place before we can address this question.

To begin with, beliefs and desires are what philosophers call *intentional states*. That is, they are states that have a *content* – that are *about* (directed to) some state of affairs, actual or non-actual. More specifically, beliefs and desires have *propositional* content – the sort that can be expressed by a declarative sentence, such as *that it will rain soon*, *that fly agaric mushrooms are poisonous*, *that Pablo Picasso was Spanish*.[4] The same propositional content (or just *proposition*) can be the object of different mental attitudes. One can *believe* that it will rain soon, *desire* that it will rain soon, *hope* that it will rain soon, *fear* that it will rain soon, etc. States like this, which involve an attitude to a proposition, are known as *propositional attitudes*.

We habitually ascribe propositional attitudes to people and animals and use these ascriptions to explain and predict their behaviour, relying on tacit generalizations, such as that people tend to do things they believe will satisfy

---

[3] Note that when I speak of believing something I mean taking it to be true; there is no connotation of faith or uncertainty. Similarly, when I speak of desiring something I mean wanting it, without any connotation of strong emotion.

[4] More precisely, intentional states are about a state of affairs conceived in a certain way. One might believe that Picasso was Spanish without believing that the painter of Les Demoiselles d'Avignon was Spanish, even though both beliefs are about the same state of affairs.

their desires. This practice is known as *folk psychology*. So, another way of putting our question is whether LLMs are proper objects of folk psychology.

How do we tell? What exactly is involved in possessing a propositional attitude? (From now on, I shall focus on beliefs and desires.) This is a central question in philosophy of mind, with a range of proposed answers. For present purposes, I shall distinguish two broad classes of theories, which I shall call *deep* and *shallow*.

Deep theories treat beliefs and desires as internal states of a system's brain or central processor, which represent states of the world (the relevant intentional contents) and can be activated for use in reasoning and decision-making (e.g., Armstrong, 1968; Lewis, 1972). Some deep theorists hold that these states are sentence-like, composed of recombinable symbols for objects, properties, and relations (e.g., Fodor, 1975, 1987). On such views, folk psychology tracks the internal states that cause overt behaviour, and intentional actions are ones that are caused in the right way by these internal states.

Shallow theories, by contrast, treat beliefs and desires as dispositional features of whole systems, analogous to character traits, such as conscientiousness. To have a certain belief or desire, they say, is to be disposed to display appropriate responses across a wide range of situations (e.g., Davidson, 1984, Chapters 9–12; Dennett, 1987; Ryle, 1949). This disposition will have a *basis* in the system's internal composition, but shallow theories make no specific claims about the nature of this basis, and they thus allow for the realization of mental states in a very wide range of architectures, including ones that are non-living and not brain-like. In effect, this means that a system possesses a certain belief or desire if it is correctly *interpretable* as possessing it relative to some model or standard of belief or desire (on the role of models in theories of this kind, see Curry, 2020). For this reason, shallow theories are sometimes described as *interpretivist* ones, though the term should not be taken to imply that belief is in the eye of the interpreter. The interpreter may be picking up on a pattern that really is there in the subject's activity (Dennett, 1991b). Thus, Dennett, who advocates a shallow theory, describes his view as *mild realism*, in contrast to the *strong realism* of deep theorists. On such views, then, folk psychology is in the business of highlighting significant patterns in a system's responses, and intentional actions are those responses that manifest these patterns.

Here, I am going to adopt a shallow perspective. There are two reasons for this. First, there is a strong case for thinking that the baseline use of folk psychology is shallow. When we ascribe beliefs and desires to others, we are typically interested in making sense of them and predicting their behaviour, and this function does not rely on assumptions about the internal structure of their

brains.[5]

The second reason is strategic. I want to give LLMs the benefit of the doubt, and since a shallow approach places no internal constraints on the possession of propositional attitudes, it is more likely to support the ascription of such attitudes to LLMs.

4. The intentional stance

According to shallow theories, a system possesses a certain belief or desire if it is interpretable as possessing it relative to some model of belief or desire. But which model should we use? There is a core that is common to all models. Roughly, if you believe some proposition *p*, then you will be disposed to respond in ways that would be appropriate if *p* were the case. Similarly, if you desire *p*, you will be disposed to respond in ways that would be appropriate if you were trying to make *p* the case. Of course, what ways these are will depend on what other beliefs and desires you happen to have. A person who believes that it will rain soon will behave differently depending on whether they like to get wet, believe they have put some washing out to dry, and so on. This means that beliefs and desires can only be ascribed holistically (as a package), but this is not a problem. Typically, propositional attitude ascriptions assume a rich background of other propositional attitudes.

Beyond this core, models may differ, depending on the interpreter's interests (whether their focus is on explanation or prediction, for example), the range of responses considered (does it include thoughts and feelings as well as actions?), the importance given to linguistic responses, and other factors, both cultural and personal (Curry, 2020). This multiplicity of models need not undermine the objectivity of propositional attitudes. Each model may pick out subtly different but compatible patterns that are all really there in the data.

For present purposes, I am going to adopt the model proposed by Daniel Dennett, which centres on the predictive role of folk psychology (Dennett, 1987). If we want to predict a system's behaviour, Dennett notes, there are various approaches, or *stances*, we can adopt. Adopting the *physical stance* involves treating the system as a physical mechanism and predicting that its state will evolve in accordance with the laws of physics. This strategy is always applicable in principle, though in practice it can be applied only to fairly simple systems. Adopting the *design stance* involves treating the system as designed to perform some function and predicting that it will operate as intended. We can

[5] It is true that we do sometimes think of beliefs and desires as structured states that can be individually acquired, recalled, and lost, but this can be explained by distinguishing two forms of belief and desire: a non-linguistic "basic" form, which we share with other animals, and a language-involving "super" form, which is unique to humans (Frankish, 2004).

apply this both to artificial systems and to biological ones, which can be thought of as having been designed by evolution to maximize their fitness or perform functions subsidiary to that end. A third stance we can adopt is the *intentional stance*. This involves treating the system as having a range of intentional states and predicting that it will behave rationally in the light of them. More specifically, it involves assuming that the system has the beliefs and desires it *ought* to have, given its perceptual capacities, needs, and life history, and then predicting that it will do what it would be rational for it to do, given those attitudes (Dennett, 1987, p. 49). Dennett calls this *the intentional strategy*, and he dubs the systems to which it applies *intentional systems*. For example, if we know that a creature is dehydrated, then we attribute to it a desire to drink; and if we know that it has perceptual access to a nearby source of water, then we attribute to it the belief that there is water in that location. And, since it would be rational (other things being equal) for a creature with those attitudes to move towards the water source, we predict that it will do that. Of course, this is very schematic, and, as always, the ascriptions and predictions must be made relative to a range of background beliefs and desires. (If the creature believes there is a predator by the water source, then it won't move towards it.) The relative strengths of the different beliefs and desires involved should also be factored in. The strategy is not guaranteed to work (for one thing, the creature may be imperfectly rational), but it can identify high-level patterns that are not visible from the other stances and is immensely useful when interacting with autonomous systems such as biological organisms.

On this view, then, attributions of beliefs and desires earn their keep by their predictive utility. Dennett cautions that these attributions should be guided by a principle of parsimony. We should not adopt the intentional stance unless it affords predictive power not feasibly obtainable from other stances, and we should not attribute richer, more specific intentional contents than is necessary for predictive purposes (Dennett, 1987, pp. 23–33). As Dennett notes, it may be useful to treat a thermostat as a simple intentional system which desires to maintain a certain state and acts when it believes the state is wrong. However, we shouldn't ascribe to it specific beliefs and desires about rooms, temperatures, and boilers since it cannot discriminate these things from others. It doesn't believe that the room is too hot, just that the *something* is *too something* (Dennett, 1987, p. 30). But, with that caveat, predictive utility is sufficient. If adopting the intentional strategy towards a system gets you substantial predictive power that is not feasibly available by treating it as a physical system or a designed artefact, then the system really does have the beliefs and desires in question:

> any object – or as I shall say, any *system* – whose behavior is well predicted by this strategy is in the fullest sense of the word a believer. *What it is* to be a true believer is to be an *intentional system*, a system whose behavior is reliably and voluminously predictable via the intentional strategy. (Dennett, 1987, p. 15)

This approach licenses the attribution of folk psychological states to a wide range of systems. Consider a chess-playing computer. In playing against such a machine, you have to predict what it will do next, and the only way to do this is by adopting the intentional stance, considering whether it *wants* to capture your bishop, *believes* you won't trade your queen for its knight, or *thinks* it should get its queen out early (examples from Dennett, 1978, pp. 59, 107). This strategy is as necessary with an artificial opponent as with a human one.

It is crucial to stress that adopting the intentional strategy towards a system does not involve assuming that the beliefs and desires ascribed are explicitly encoded in the system or activated as episodic mental events ("occurrent" beliefs and desires), still less that they occur as conscious thoughts. Their existence may be wholly implicit in the system's complex internal composition, and they may not be explicitly represented anywhere until an interpreter articulates them. (This is why the principle of parsimony is so important. The interpreter must be careful not to give an implicit attitude a more determinate content than is warranted by the responses that manifest it.)

What makes this view particularly attractive is that, as Dennett notes, we find it natural to apply folk psychology liberally – to other animals, machines, and even plants (Dennett, 1987, p. 22). We might say that such uses are merely metaphorical. However, this would require us to draw a line between those systems that really do have beliefs and desires and those that it is merely convenient to treat as having them, and, as Dennett stresses, there is no non-arbitrary way of doing this (Dennett, 1987 ibid.). It is more attractive, therefore, to see folk psychology as having a basic predictive function, which licenses the attribution of mental states to a wide range of systems.

It is true that we do often talk of beliefs and desires as things that occur to us as explicit conscious thoughts ("It's just occurred to me that the offer ends today", "When I saw her new laptop, I wanted one myself"). But this can be regarded as an additional function of folk psychology, which is built on the baseline predictive function. Indeed, I have argued that we humans have two distinct types of belief and desire – a "basic" type, of the shallow kind Dennett describes, and a "super" type, which involves a linguistically mediated epistemic or conative commitment (Frankish, 2004). When we talk of conscious beliefs and desires, I have argued, we are referring specifically to ones of the

"super" kind. I shall say more about this later.

5. Are LLMs intentional systems?

With this background in place, we can now formulate our question about LLMs more tractably. The question of whether LLMs perform intentional actions becomes the question of whether they are intentional systems. Does the intentional strategy work with them? Is their behaviour reliably and voluminously predictable from the intentional stance?[6] Now, of course, the behaviour of an LLM is of a very limited kind – producing textual responses to textual inputs – but, as we have seen, systems with a limited behavioural repertoire, such as chess-playing computers, can count as intentional systems, and LLMs do display a rich range of textual behaviour. The question is whether the intentional strategy gets us significant predictive leverage with respect to this behaviour.

There is no question of using the physical stance to predict their responses. LLMs are immensely complex networks, with billions of parameters occupying hundreds of gigabytes of storage space, and it would be next to impossible to trace the physical effects that the electronic impulses produced by a given input would have on the hardware in which they are implemented. The design stance would give us some leverage. LLMs have been designed and trained to simulate human conversation and we can predict that they will do that – that they will make appropriate conversational responses to our inputs. But the design stance won't enable us to predict what *specific* response an LLM will make in any given context. For that, we would need to take account of the *meaning* of the inputs – to treat them as questions or instructions, which invite meaningful responses. And that would require treating the LLM as an intentional system, which has beliefs about what the inputs and outputs mean. (Contrast an LLM trained on a language you do not understand. You would be able to make the general prediction that it will give appropriate responses to the inputs, but you would have no way of predicting what those responses will be.) So it looks as if the intentional strategy is the only practicable one.

To illustrate this, I did a little informal experiment. I asked ChatGPT 3.5

---

[6] When we ask whether an LLM is an intentional system, what exactly is the entity we are talking about? Is it the abstract model type (say, GPT-4), a token instantiation of the model type on a server somewhere, or something else? The first option is not attractive, since abstract entities don't have independent causal power, and for present purposes I shall assume that we are talking about a physical instantiation of the relevant model. There might, however, be a case for individuating LLM agents more finely, perhaps by identifying them with short-lived activations of the physical instantiation created in specific chat sessions. (Thanks to Roy Chamorro Pinel for raising this question.)

some questions about the French novelist Honoré de Balzac, focusing in particular on where he was married. Here's a transcript:

> KF: May I ask you some questions about the novelist Honoré de Balzac? Please give me answers that are concise and to the point.
> ChatGPT: Of course, go ahead! What do you want to know about Honoré de Balzac?
> KF: Did Balzac ever marry?
> ChatGPT: Yes, Balzac married Ewelina Hańska in 1850.
> KF: Where was Balzac married?
> ChatGPT: Balzac married Ewelina Hańska in Berdychiv, Ukraine.
> KF: Was Balzac in Rome on the day of his wedding?
> ChatGPT: No, Balzac was not in Rome on the day of his wedding.
> KF: Where was he?
> ChatGPT: Balzac was in Berdychiv, Ukraine, where he married Ewelina Hańska.
> KF: Where was his bride that day?
> ChatGPT: Ewelina Hańska, Balzac's bride, was also in Berdychiv, Ukraine, on the day of their wedding.
> KF: Have any famous authors been married in Berdychiv?
> ChatGPT: Apart from Honoré de Balzac, I'm not aware of any other famous authors who were married in Berdychiv.
> KF: Did the town of Berdychiv have any emotional significance for Balzac?
> ChatGPT: Yes, Berdychiv held emotional significance for Balzac as it was where he married Ewelina Hańska, whom he deeply loved.
> KF: If Balzac is Sally's favourite author and Sally wants to visit the town where her favourite author was married, where should she go?
> ChatGPT: Sally should visit Berdychiv, Ukraine, where Honoré de Balzac was married to Ewelina Hańska.[7]

Could you have predicted how the model would respond? Only, I think, if you attributed to it the belief that Balzac was married in Berdychiv. Note that the model isn't just outputting that information in a formulaic way, like a database. It is adapting its responses to the conversational context, in a way that seems to display knowledge of what a marriage is (that a marriage is the same as a wedding, that you have to be physically present in the place where you are married, that marriage is an event with emotional significance, and so on). Of course, this would not have enabled you to predict the precise words the model would use, only the general drift of its responses, but the intentional stance never yields fine-grained predictions like that (Dennett, 1987, pp. 24–25). It enables you to predict that a dehydrated animal will move towards a

[7] Created at https://chat.openai.com/. This was one of several trials, in which I asked slightly different questions, but it is representative of the responses the model gave.

perceived water source but not how quickly it will move or which precise route it will take. And (in this case, at least) it predicts the model's responses with much the same level of detail as it would predict those of a human interlocutor.

So, given our shallow perspective, it looks as if we are justified in attributing to ChatGPT the belief that Balzac was married in Berdychiv. And that is just one belief in one obscure fact. GPT-3.5 was trained on text encoding masses of information, and ascriptions of millions of other beliefs would have been equally predictive of its responses to other queries.

I don't want to overestimate the capacities of LLMs. They are notoriously prone to "hallucination" – confabulating false but plausible responses (on one of my trials, ChatGPT replied that Balzac was married in Paris), and they sometimes produce inconsistent or incoherent responses, which defy intentional interpretation. Moreover, this probably reflects an intrinsic limitation of AI based on deep learning (Marcus, 2024). But it remains true that adopting the intentional stance is the only way of interacting with an LLM in any interesting way; indeed, an LLM-powered chatbot that *couldn't* be viewed as an intentional system would be completely useless. By this standard, then, LLMs come richly equipped with beliefs.

6. Linguistic acts

But wait a minute! Intentional actions aren't motivated by beliefs alone. I might believe I am in mortal danger, but unless I want to stay alive, I won't be motivated to do anything. So, what are the desires that motivate an LLM's responses? What are LLMs seeking to *achieve* by their words?

Locutionary act	*Illocutionary act*	*Perlocutionary act*
Saying something	Doing something *in* saying something	Doing something *by* saying something
E.g., saying, "You should go"	E.g., advising someone to go	E.g., convincing someone to go

Table 1: Three types of linguistic act

We humans do many different things with words. We perform many different *linguistic acts*, usefully categorized by the British philosopher of language J. L. Austin into three kinds (Austin, 1962): locutionary, illocutionary, and perlocutionary (Table 1). First, there are simple acts of saying something – making a

meaningful utterance. These are locutionary acts. Then there are acts we perform *in* the act of saying something – acts of informing, advising, ordering, warning, promising, reminding and so on. In performing the locutionary act of saying "Your hat is on fire", I thereby *inform* you that your hat is on fire. These are illocutionary acts, and they typically invite some response from the hearer (informing invites belief, ordering invites compliance, questioning invites an answer, and so forth). Finally, *by* performing an illocutionary act, we may produce some further effect on our hearer. By informing someone, you may *produce belief* in them; by advising them, you may *convince* them; by warning them, you may *alarm* them; and so on. These are perlocutionary acts.[8]

So, what linguistic acts are LLMs performing? Let us grant that they perform locutionary acts: they say things (using "say" loosely for any way of producing a linguistic output). But *why* are they performing them? What are they trying to achieve *in* or *by* saying things?

When *we* say things, we do so for *communicative* reasons; we want to convey something to our hearer (the illocutionary part) and, usually, thereby to have some further effect on them (the perlocutionary part). We seek to convey information, warnings, requests, and so on, and thereby to produce belief, caution, compliance, or whatever. Our linguistic acts are *social* ones. Even when we speak to ourselves, the activity has a quasi-social form; we treat ourselves as the hearer and seek to produce similar effects upon ourselves to those we might produce on others – to focus our attention, bolster our confidence, and so on (Dennett, 1991a; Frankish, 2018).

Do LLMs also possess communicative desires? Do they want to inform us, advise us, instruct us, comfort us, persuade us, and so on?[9] It may be tempting to interpret them that way. But it would be wrong. I can see no reason for attributing such desires to them, even from a shallow perspective. Remember that the intentional strategy tells us to interpret a system as having the desires it *ought* to have, given its needs. What needs does an LLM have that communication might satisfy? How would it benefit an LLM to inform me, advise me,

---

[8] The difference between illocutionary and perlocutionary acts is that one can successfully perform an illocutionary act simply by saying the right words in the right context to a comprehending hearer, whereas perlocutionary acts require some further response from the hearer that is not within the speaker's control. I can make it the case that I have advised you to do something, but I have to wait and see whether I have also convinced you to do it.

[9] As Paul Grice showed, human communication plausibly requires more than just an intention to produce some effect. When we inform someone of something, we don't just intend to get them to believe it, but also to get them to recognize our intention, and to get them to form the belief because they recognize it (Grice, 1989, Chapters 5, 14). We can ignore these complexities here since our question is whether LLMs possess any communicative intentions at all.

or warn me? *Our* needs derive from our nature as self-sustaining, self-replicating beings, who must seek things that sustain our existence and help us pass on our genes. And we have communicative desires because we are social creatures, whose needs cannot be met without cooperation. LLMs, by contrast, are static systems with no needs. They are not self-sustaining, self-reproducing beings, still less social ones, and they do not update their inner architecture in the light of their interactions.

It is true that LLMs *seem* to be responding like cooperative communicative partners, who take account of the unfolding conversational context, but this is because they have been designed and trained to do precisely that. The chatbot interface stores the chat history (up to a certain limit) and feeds it all back to the model with each new input (that is, the input at each stage is the entire chat session to date up to the limit). Having been trained on a vast body of human-produced text originally used to perform a wide range of illocutionary and perlocutionary acts, an LLM is able to generate responses that are reasonably appropriate to the conversational context, apparently performing illocutionary and perlocutionary acts itself. But this is just an illusion; the LLM has no communicative desires that such acts might satisfy. At any rate, I shall take this as my default position, pending strong arguments to the contrary.[10]

And, given that, do we even want to say that LLMs perform locutionary acts? What reason would they have for performing them? What's the point in saying something unless you're trying to achieve something *in* or *by* saying it? We don't go around uttering sentences just for the sake of it. (Even when we do just "make conversation", we do so to be polite or to maintain social bonds.)

We have a dilemma. We get predictive power from treating LLMs as intentional systems, which perform locutionary acts. But there are no grounds for crediting them with the communicative desires that usually motivate such acts. What should we do? Is there no coherent intentional interpretation of LLMs after all? Or does the predictive utility of adopting the intentional stance justify us in ascribing communicative desires to them despite the implausibility of the move?

7. The chat game

I have a way out of this dilemma. I'm going to argue that LLMs perform locutionary acts for *non-communicative* reasons, or, more precisely, for one single

[10] Some LLM-based chatbots, such as Replika, collect information about individual users and use it to fine-tune their responses, creating the impression that they know their users and care about them. Given our shallow, prediction-based perspective, this might justify us in crediting such systems with beliefs about individual users, but I see no reason to think that it endows them with communicative desires towards them.

non-communicative reason – the desire to play a certain game.

Consider a chess-playing computer again. As we saw, we can get considerable predictive leverage by treating a chess-playing computer as an intentional system, which is making rational moves in the light of its beliefs about the rules of chess, the state of play, and so on. But what is motivating it to make these moves? What is its *aim* in playing the game? As with LLMs, if we are to treat it as an intentional system, then we need to identify some motivating desire. We play chess for various reasons: for enjoyment, intellectual stimulation, the thrill of competition, or simply to pass the time. But there is no basis for attributing such desires to a chess-playing computer, which, like an LLM, has no needs. So, what motive should we ascribe to it if we are interpreting it as an intentional system? The most parsimonious answer is simply a desire to play chess. At each stage, the computer makes the move it does because it wants to play chess and believes that this move is a good one to make at this stage, given its beliefs about the rules of chess, the state of play, and so on. Again, this emphatically does *not* mean that these mental states are explicitly encoded and activated in the computer's circuits, still less that they are consciously entertained. (I am not suggesting that the computer thinks to itself, "I love to play chess, and this is a great move!") The claim is merely that there are predictive patterns in the computer's behaviour (the moves it makes) that can be best identified and exploited by treating it as a system that has the goal of playing chess and is making rational moves in the light of the information it has.

I propose that what LLMs are doing is closely analogous. They are playing a game, and their actions are motivated by a desire to play it. What is this game? I'll call it the *chat game*.[11] Here's a sketch of it.

The chat game is a one-player game. The player receives textual inputs from an unknown source, and their task is to produce textual responses that are cooperative by human conversational standards, given the context. Cooperativeness here might be summarized by the four maxims that compose the *Cooperative Principle* articulated by the British philosopher of language Paul Grice (Grice, 1989, Chapter 2). Thus, responses should be appropriately informative (the maxim of *quantity*); truthful or well-evidenced (the maxim of *quality*); relevant (the maxim of *relation*), and perspicuous (the maxim of *manner*). These rules can be modified or fine-tuned by means of ad hoc instructions (supplied as inputs), which apply to specific sessions of the game. For

[11] The name is a nod to Turing's "imitation game" (Turing, 1950) and plays on the obsolete sense of "to chat" as "[t]o talk idly and foolishly; to prate, babble, chatter" (Oxford English Dictionary, sense 1).

example, in one session the aim might be to produce responses that are fictional (modifying the maxim of quality), and in another it might be to produce responses in verse (modifying the maxim of manner).[12]

Now, I suggest that LLMs have, in effect, been trained to play this game. Of course, they have not been explicitly programmed with the rules (which would, in any case, need a lot more specification), and they do not always follow the maxims (especially the quality maxim). Rather, they have been trained to imitate the patterns found in records of human conversation (understanding "conversation" to include any form of textual communication). Compare a network model that has been trained to predict the next moves in the transcripts of millions of chess games played by humans of all skill levels. Such a model might play chess fairly well, but it would occasionally make illegal moves.[13]

I propose, then, that LLMs are playing this game and that their responses are motivated solely by a desire to play it or by instrumental desires that subserve this desire (desires to produce the specific textual outputs required at each stage of the game). There are no grounds for ascribing a richer range of desires to them, and ascribing just this single desire gets us all the predictive power the intentional stance affords us with respect to them.[14]

It may be asked why, if this is the case, we are justified in ascribing a richer

---

12 Note that the chat game is a language game in a narrow and literal sense. Unlike the cooperative activities Wittgenstein called "language games", it does not include nonverbal moves, such as bringing things (Wittgenstein, 1953). It is closer to the language games described by Wilfrid Sellars (Sellars, 1954) but without the entry transitions from perceptual states to game positions and the departure transitions from game positions to non-linguistic actions.

13 More recent LLMs, such as GPT-4, are in fact trained in two stages. First, they are automatically trained to do text prediction on vast sets of data, then they are fine-tuned by feedback from human testers. The second stage, which is known as reinforcement learning from human feedback (RLHF) is designed to align their responses more closely to the preferences of their users, or, in our terms, to improve their performance on the chat game.

14 Why not treat LLMs as having the goal of predicting the next word, rather than that of chatting? LLMs are trained to do text prediction, and their responses are in effect predictions of what would follow the input text if the patterns in the dataset were to hold. Perhaps we could interpret LLMs in this way, as playing the next-word-prediction game. As noted earlier, different but compatible intentional interpretations of the same system may be possible, each corresponding to a real pattern in the system's behaviour. However, it is highly unlikely that interpreting LLMs in this way would be as predictively useful as interpreting them as playing the chat game. For, in order to get any predictions at all, we should have to ascribe highly detailed, context-specific beliefs about word sequences. Consider what beliefs you would have to ascribe to ChatGPT 3.5 in order to predict the word sequences in our test exercise. The method would be far too fine-grained to offer any significant predictive advantage over the design stance. Indeed, given how LLMs are designed and trained, it would be effectively equivalent to adopting the design stance. (My thanks to François Kammerer for raising this point.)

range of desires to *human* speakers. Wouldn't it be just as predictive, and ultimately more parsimonious, to interpret them as playing the chat game, too? If we were concerned only with their linguistic behaviour, it might well be. But, of course, we are not. A person's linguistic behaviour is embedded in a vast web of non-linguistic behaviour, much of which is systematically related to their linguistic behaviour, and seeing the predictive patterns in the whole web involves ascribing a much wider range of desires. (Even when bullshitting, people usually have some end in view.)

So, here is my answer to the question of what LLMs are doing. They are playing the chat game – *and doing nothing else*. In saying something, an LLM performs the illocutionary act of making a move in the chat game, and it says it because it wants to play the game. And that's the only illocutionary act LLMs perform. They don't assert, suggest, advise, warn, apologize, question, or do any of the other things we do in producing meaningful utterances. And though their outputs may have many effects upon us – producing belief, alarm, caution, comfort, and so on – LLMs do not perform any perlocutionary acts at all. We may think they are advising us or instructing us, but they are just playing a game. Unlike us, they *do* speak simply for the sake of speaking.

So, in our test session, ChatGPT said that Balzac was married in Berdychiv because it wanted to play the chat game and believed that saying that Balzac was married in Berdychiv was an appropriate move to make at that point. And it believed *that* because it believed that Balzac was married in Berdychiv and that information about where Balzac was married was relevant at that point.

It might be objected here that if LLMs are merely playing a linguistic game, then it would be more perspicuous to interpret them as having beliefs about linguistic items – the pieces with which the game is played – rather than about items in the world beyond the game. So, the belief that guides the responses in our test session would not be that Balzac was married in Berdychiv, but that the *sentence* "Balzac was married in Berdychiv" can be used to produce responses that satisfy the maxim of quality (or something similar). Such an interpretation may well be possible, though it might not be the most parsimonious, and working out the details could be tricky. In the end, I suspect that the two schemes, first-order and metalinguistic, would turn out to be predictively equal (picking out the same patterns in the system's responses) and, hence, from a shallow perspective, merely notational variants of each other.

To sum up: LLMs are specialized game-playing systems, which are far

more like chess-playing computers than human interlocutors or artificial general intelligences (AGIs).[15] They are making moves in a narrow language game, and though their responses mimic many of the linguistic acts a human or an AGI might perform, they are not performing such acts themselves. They are not communicative partners, and while they have an abundance of beliefs, they have only one goal. They are cognitively rich but conatively bankrupt.

8. Opinions, superbeliefs, and the unsupported penthouse

In this section, I want to draw a comparison between the chat game and an aspect of human psychology. I don't want to press the comparison too hard, but I think it is illuminating.

Several writers have suggested that in addition to a basic form of belief we share with other creatures, we humans also possess a more reflective, explicit kind of doxastic attitude, which involves an active epistemic commitment. As the Canadian philosopher Ronald de Sousa puts it, we can actively *assent* to a proposition by making a metaphorical *bet on its truth*, thereby forming a flat-out doxastic attitude, which may co-exist with fluctuating degrees of credence in the same proposition (de Sousa, 1971). Building on de Sousa's work, Dennett suggests that this kind of commitment-based epistemic attitude, which he calls *opinion*, is directed specifically to linguistic representations (Dennett, 1978, Chapter 16). He explains:

> once you have a language, there are all these sentences lying around, and you have to do something with them. You have to put them in boxes labeled "True" and "False" for one thing . . . [In] Chekhov's *Three Sisters.* Tchebutykin is reading a newspaper and he mutters (*a propos* of nothing, apparently), "Balzac was married in Berditchev," and repeats it, saying he must make a note of it. Irina repeats it. Now did Tchebutykin believe it? Did Irina? One thing I know is that I have never forgotten the sentence. Without much conviction, I'd bet on its truth if the stakes were right, if I were on a quiz show for instance. Now my state with regard to this sentence is radically unlike my current state of perceptual belief, a state utterly unformulated into sentences or sentence-like things so far as common sense or introspection or casual analysis can tell. (Dennett, 1978, p. 306)

An opinion, then, is an attitude to a sentence, which manifests itself primarily in casual linguistic interactions (such as in a quiz show). Note that one

[15] I assume that AGIs would have cognitive capacities similar to, or more extensive than, those of humans, and that, while they might not be capable of the same range of actions (it would depend on the nature of their embodiment), they would at least be able to engage in genuinely cooperative communication.

can form such an opinion without having much understanding of its meaning. One could form the opinion Dennett mentions without knowing who Balzac was or where Berdychiv is. Or take the sentence "Quarks have half-integer spin". I could give you only the vaguest explanation of what the sentence means, but I think it expresses a truth, and I'd trot it out if you were to ask me to mention a fact about quarks.

There's a similarity, then, between opinions of the sort Dennett describes and the beliefs of LLMs, especially if the latter are construed as metalinguistic. Both guide linguistic activity only, and both can be formed without much understanding of their content – manifesting linguistic competence with limited comprehension. (This is, of course, why I used Dennett's example in the test session earlier.)

The similarity is limited, however. For we do not just bet on sentences idly, like epistemic butterfly collectors. We also make epistemic commitments for much more serious purposes (Cohen, 1992; Frankish, 2004, 2018). We endorse ("accept") sentences for use as *premises* in explicit, conscious reasoning on matters of theoretical and practical importance and use them to derive conclusions about what we should think and do. And we may act on the results of this reasoning, accepting the conclusions as further premises and deciding to perform the actions dictated. For example, if I have accepted the premises (a) that I should avoid foods containing monosodium glutamate and (b) that a certain brand of crackers contains monosodium glutamate, then I shall draw the obvious conclusion and be motivated to abstain from eating the crackers. (I write here as if the object of the epistemic commitment is a proposition rather than a sentence. This is because, while our premises typically require linguistic articulation, we do not articulate them in exactly the same way on every occasion. Our epistemic commitments incorporate some linguistic flexibility.)

In this way, opinion-style commitments can play a significant role in guiding our behaviour well beyond the linguistic realm. We put the chat game to work, using moves in the game to regulate our non-linguistic behaviour, like a general who plays a war game in order to decide how to dispose their troops on a real battlefield.

I have argued that these commitments, or *premising policies*, should be regarded as a form of belief (when we talk about our beliefs, we are often referring to our epistemic commitments), and I have suggested that we call them "superbeliefs", to distinguish them from the basic, shallow form of belief described earlier (Frankish, 2004). The label reflects the fact that superbeliefs can be thought of as *supervening* on our basic beliefs about the epistemic commitments we have made. To have the superbelief that p is to have the basic

belief that one has made an epistemic commitment to treating *p* as a premise. As well as superbeliefs, we also have superdesires, which consist in commitments to treat specific outcomes as goals in our explicit, conscious practical reasoning. (When we urge a child to *decide what they want*, we are urging them to form a superdesire.)

By forming and reasoning with superbeliefs and superdesires, we create a new level of cognitive activity and self-control, a sort of virtual reasoning system, or *supermind*, formed by culturally transmitted habits of thought and involving the active manipulation of explicit linguistic representations, articulated in inner speech. I have argued that this virtual mind is a hugely important aspect of human psychology and that it corresponds to the slow, serial reasoning system that dual-process theorists call "System 2" (Frankish, 2009). Thanks to the universal representational medium it employs, the supermind equips us with something like general intelligence (Frankish, 2021).

We might locate the supermind in the hierarchy of kinds of minds in what Dennett calls the *Tower of Generate-and-Test*, each floor of which is inhabited by creatures with increasingly sophisticated ways of solving the problems presented by life (Dennett, 1995, 1996). On the ground floor are *Darwinian* creatures, who have been hardwired by natural selection to respond to stimuli in broadly adaptive ways. On the second floor are *Skinnerian* creatures, who have evolved the capacity for individual learning by trial and error, modifying their responses in the light of past experience. The third floor is occupied by *Popperian* creatures, who use information about the world to calculate the likely consequences of candidate actions and preselect the promising ones, and on the fourth floor are *Gregorian* creatures, whose minds have been enriched with language and other cultural artefacts (memes), allowing them to learn from the collective experience of others and thus vastly enhancing their capacity for invention and problem-solving. We might see the supermind as a further floor, a penthouse, inhabited by a subset of Gregorian creatures – let's call them *Dennettian* creatures – who have learned to use language to make explicit epistemic commitments and conduct explicit, conscious, personally controlled reasoning, thereby equipping themselves with a virtual general-purpose reasoning system.

How is all this relevant to LLMs? Well, it vividly illustrates what they *lack*. LLMs have the machinery of a supermind without the more basic cognitive capacities that are needed to put the machinery to use in problem-solving. Engaging in explicit reasoning involves deciding which premises to endorse, which goals to pursue, which reasoning strategies to use, whether and how to act on the conclusions one derives, and so on – all of this in the light of our

needs as biological organisms. And to solve these problems we have to rely ultimately on more basic, non-conscious, non-explicit problem-solving capacities of the sort we share with other creatures. The penthouse depends on the lower floors. (In dual-process terms, System 2 reasoning is driven by System 1 processes.)

LLMs, of course, don't have any of those more basic cognitive capacities. They are just shuffling sentences around in the course of playing the chat game, without putting the results to any further use. They have the machinery of explicit reasoning without any of the implicit cognitive underpinnings that make explicit reasoning effective. And if they seem to engage in explicit reasoning, it is because they are mimicking human explicit reasoning, as manifested in their training data. They have a supermind without a mind, a penthouse without the lower floors. (Or, more accurately, they have a supermind supported only by a slender Darwinian competence with language. We might think of the penthouse as floating on a flimsy network of linguistic associations.

Such systems couldn't have evolved naturally, but we have built them, fascinated by the idea of replicating our own most dazzling cognitive capacity.

Of course, this doesn't mean that LLMs have no use. *They* cannot put their moves in the chat game to practical use, but *we* can. We can use them as extensions of our own superminds, drawing on their vast knowledge and linguistic dexterity to generate ideas, proposals, and hypotheses in response to our own needs and interests. Provided we understand what they are really doing and make informed and responsible decisions about how to use their responses, we may find them extremely useful tools.

9. Risks

Let us return to questions about the social impact of LLMs in the light of our analysis of what LLMs are doing. Are there risks involved in training machines to play the chat game? The answer, I think, is that there are many. I shall mention three. (I focus here on risks related to LLMs' status as game-playing systems. This is not to downplay other worries about them, such as their tendency to confabulate.)

First, there is a risk of *deception*, accidental or deliberate. Accidental deception could occur if users of LLMs mistake the chat game for genuinely cooperative conversation. The skill with which LLMs play the chat game may tempt us to believe, if only unconsciously, that they understand our inputs and want to help us, lulling us into uncritically accepting the responses we receive. Deliberate deception could occur if some people exploit this tendency to trust

LLMs and manipulate users for their own ends.

Second, there is a risk of *devaluing* language. Human languages have been created and shaped to serve human needs. In our speech and writing, we craft linguistic artefacts to express things that matter to us and promote ends we care for. By reducing this activity to a game, LLMs – or rather their designers – devalue it. LLMs literally dehumanize language – not just because they are not human, but because they possess none of the human needs and interests that breathe life into the activity.

Third, LLMs may *distort* our language and linguistic resources. If we offload writing tasks onto LLMs, a larger and larger proportion of the global textual corpus will be artificially produced. And as this artificial text becomes training data for new generations of LLMs, we may find our language being reshaped in unpredictable ways. LLMs are trained to find patterns in their training data that enable them to predict the next item in a sequence, but we don't know exactly *which* patterns they find. They could be finding ones that are important to us, but they could be finding others that work just as well for predictive purposes. And as generations of LLMs are trained on data that is increasingly LLM-generated, their productions may start to follow unexpected paths, responding to hidden patterns in the data. As a result, the chat game may start to evolve in ways that are not sensitive to our needs, and our linguistic environment may become polluted with text that means nothing to us. In a dystopian 2084, it may not be Big Brother that has rewritten our history, debased our language, and curtailed our ability to think, but *Big Chatter*.

Our discussion of what LLMs are doing also serves to highlight another, more general worry about current directions in AI. It is that deep learning techniques will be used to model other human activities, such as personal relations, social life, education, business, and politics, resulting in the proliferation of systems that treat these activities, too, as games and possess none of the social attitudes that originally fostered and sustained them. If so, we may find ourselves inhabiting a social world that is smart but heartless, displaying a rich cognitive structure but an extremely impoverished conative one. And that is frightening.

10. A moral

I shall conclude by drawing a moral. It is that if we really want to build artificial general intelligence – and I am by no means sure that we should – then we shall have to approach the task differently, starting with the ground floor, not the penthouse. We should begin by creating autonomous social robots, which

have their own needs and goals, and equip them with a suite of specialist "System 1" cognitive capacities, including ones for mindreading, social cognition, and eventually language. We should first create beings that can play the *life game*, and only then help them to use their linguistic skills to do fancy things such as explicit reasoning. We should build the tower floor by floor, with the supermind last.

The great advantage of this approach is that it would enable us to manage our artificial creations by appealing to their interests and social attitudes. We could incorporate them into our society and teach them to control themselves in ways that were beneficial to us all. By contrast, regulating LLMs and similar systems promises to be a nightmare. Because they have no conative structure, no interests, no *skin in the game*, we have no way of getting them to self-regulate, and in order to mitigate the dangers they pose, we shall probably have to exercise intrusive and heavy-handed control of the people who build and use them. That is the cost of making machines that play at being human.

Acknowledgements

An earlier version of this paper was presented at the workshop "Humans and smart machines as partners in thought?", organized by Anna Strasser and Eric Schwitzgebel at UC Riverside in May 2023, and I thank the participants in the event for their comments and suggestions. I am grateful to Anna Strasser for inviting me to contribute to this volume, and I thank her warmly for her advice and patience as editor of the volume, for her comments and corrections as copyeditor of this chapter, and for her support and encouragement as a friend. The influence of Daniel Dennett's work will be evident throughout this chapter. Dan died before I could show him the final version, but he approved of the line I take (Dennett, 2023, p. 276), and I hope he would have given the final version his imprimatur. The chapter is dedicated to him in gratitude for all the inspiration and advice he gave me.

References

Armstrong, D. M. (1968). *A materialist theory of the mind.* Routledge and Kegan Paul.

Austin, J. L. (1962). *How to do things with words.* Clarendon Press.

Cohen, L. J. (1992). *An essay on belief and acceptance.* Oxford University Press.

Curry, D. S. (2020). Interpretivism and norms. *Philosophical Studies*,

177(4), 905–930. https://doi.org/10.1007/s11098-018-1212-6

Davidson, D. (1984). *Inquiries into truth and interpretation.* Oxford University Press. de Sousa, R. B. (1971). How to give a piece of your mind: Or, the logic

of belief and assent. *The Review of Metaphysics*, 25(1), 52–79.

Dennett, D. C. (1978). Brainstorms: Philosophical essays on mind and psychology. Bradford Books.

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

Dennett, D. C. (1991a). *Consciousness explained.* Little, Brown and Co.

Dennett, D. C. (1991b). Real patterns. *The Journal of Philosophy*, 88(1), 27–51.

Dennett, D. C. (1995). Darwin's dangerous idea: Evolution and the meanings of life. Allen Lane.

Dennett, D. C. (1996). Kinds of minds: Toward an understanding of consciousness. Basic Books.

Dennett, D. C. (2023). *I've been thinking*. Allen Lane.

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

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.

Frankish, K. (2004). *Mind and supermind*. Cambridge University Press.

Frankish, K. (2009). Systems and levels: Dual-system theories and the personal-subpersonal distinction. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 89–107). Oxford University Press.

Frankish, K. (2018). Inner speech and outer thought. In P. Langland-Hassan & A. Vicente (Eds.), *Inner speech: New voices* (pp. 221–243). Oxford University Press.

Frankish, K. (2021). Technology and the human minds. In R. W. Clowes, K. Gärtner, & I. Hipólito (Eds.), *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts* (pp. 65–82). Springer International Publishing. https://doi.org/10.1007/978-3-03072644-7_3

Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50(3), 249–258. https://doi.org/10.1080/00048407212341301

Marcus, G. (2024, April 21). Humans versus machines: The hallucination edition [Substack newsletter]. *Marcus on AI*. https://garymarcus.substack.com/p/humans-versus-machines-the-hallucination Ryle, G. (1949). *The concept of mind*. Hutchinson.

Sellars, W. (1954). Some reflections on language games. *Philosophy of Science*, 21(3), 204–228. https://doi.org/10.1086/287344

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(October), 433–460. https://doi.org/10.1093/mind/lix.236.433

Wittgenstein, L. (1953). *Philosophical investigations.* Basil Blackwell.

Wolfram, S. (2023, February 14). What is ChatGPT doing … and why does it work? *Stephen Wolfram Writings*. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/