**Artificial Neural Networks and the Intentional Stance**

Thomas Raleigh, University of Luxembourg, thomas.raleigh@uni.lu

**INTRODUCTION:**

Deep Neural Networks (DNNs) can now perform many tasks at human or greater-than-human levels of capability.

But, we don’t really understand *how* they are performing the functions they perform – hence DNNs are called ‘black boxes’ or ‘opaque’.

The field of Explainable AI (XAI) aims to create simplified models of a DNN so as to provide a humanly comprehensible explanation of how it transforms inputs to outputs.

There has been much discussion as to whether/when XAI models really can provide genuine explanation/understanding

**THE INTENTIONAL STANCE:**

Dennett (1971, 1987) famously drew a distinction between 3 explanatory strategies:

* The Physical Stance – treat the explanandum as physical stuff subject to physical forces and laws.
* The Design Stance – treat the explanandum as having a function that it is designed to perform.
* The Intentional Stance – treat the explanandum as an approximately rational agent, whose aims and plans and beliefs etc. explain its behaviour.

“The best chess-playing computers these days are practically inaccessible to prediction from either the design stance or the physical stance; they have become too complex for even their own designers to view from the design stance. A man's best hope of defeating such a machine in a chess match is to predict its responses by figuring out as best he can what the best or most rational move would be, given the rules and goals of chess. That is, one assumes not only (1) that the machine will function as designed, but (2) that the design is optimal as well, that the computer will "choose" the most rational move. Predictions made on these assumptions may well fail if either assumption proves unwarranted in the particular case, but still this means of prediction may impress us as the most fruitful one to adopt in dealing with a particular system. Put another way, when one can no longer hope to beat the machine by utilizing one's knowledge of physics or programming to anticipate its responses, one may still be able to avoid defeat by treating the machine rather like an intelligent human opponent.” (Dennett, 1971, 89)

Actually Dennett endorsed both of the following theses:

(1) Adopting the intentional stance is, sometimes, a legitimate way of predicting and explaining the behaviour of a complex system – i.e. despite the simplification/idealising involved it can provide genuine (though partial) understanding of the system.

(2) That a system can be richly and usefully interpreted from the intentional stance is what ultimately grounds or explains the attribution of representational contents/states to the system.

But it seems that one could endorse (1) without necessarily endorsing (2).

Sometimes Dennett’s opponents characterize him as having an anti-realist or instrumentalist view about beliefs and desires – i.e. folk psychology is *merely* a useful instrumental tool for predicting behaviour. But in fact his view is more nuanced and more realist than that – e.g. see his classic paper ‘Real Patterns’ (1991).

I will assume that Dennett is correct about thesis (1) – I assume that the intentional stance can sometimes provide genuine, though partial, explanation/understanding. (I will remain neutral about (2) for today.)

**ZERILLI:**

‘…we can adopt an intentional stance towards ML systems – i.e. to show in what sense it could pay dividends to do so. This is just to make the claim that ML systems are rational (much like Dennett’s example- in-chief, the chess-playing computer programme).’ (Zerilli, 2022, 11)

‘if intentional explanations are considered good enough for assessing human decisions, they should be considered prima facie good enough for assessing automated decisions.’ (ibid, 8)

‘there is no basis for thinking that the “reasons” of a supervised ML system would be even less faithful guides to its behaviour than human reasons are to human behaviour.’ (ibid, 10)

‘…a significant body of work in XAI… aims to explain ML systems by reducing their operations to a form that is amenable to belief-desire representation.’ (ibid, 2)

Let’s distinguish 3 theses from the Zerilli quotes:

(i) (Some) DNNs can (sometimes) be usefully explained/understood using intentional stance.

(ii) Humans are just as opaque as DNNs and so intentional stance explanations (assuming they are predictively successful) should be no more or less genuine/legitimate for one than for the other.

(iii) Many existing XAI technologies are in the business of supplying intentional stance explanations of DNNs.

Note: thesis (i) is a fairly weak possibility claim.

Note also: one could endorse (i) without endorsing (ii) or (iii).

I think (i) is correct, but I will argue that the specific kinds of DNNs that Zerilli focuses on are *not* in fact usefully interpreted/explained from the intentional stance – because here the intentional stance does not bring any explanatory advantage over taking the design stance. However, I will suggest that there are other varieties of DNN which *are* usefully understood from the intentional stance.

[I will remain neutral about (ii) and (iii) for today.]

The kinds of DNNs that Zerilli focuses on are standard feed-forward networks trained by supervised gradient descent to perform classification or prediction tasks.

E.g. networks for image classification, networks for predicting the risk of recidivism amongst prisoners, networks for diagnosing cancer, networks for deciding whether to approve a bank loan.

Recall the constraint that the intentional stance must offer considerable advantage over the design stance. Or as Dennett puts it: ‘your problem of predicting and interpreting their behaviour is made vastly easier than it would be if you tried to use the physical or the design stance.’ (Dennett 1971, 340)

Think also of how in the example of a chess-playing computer, we can usefully attribute quite complex instrumental desires/intentions that sub-serve the overall aim of winning the game.

But when we consider DNNs trained by supervised learning to perform classification/prediction tasks, their behaviour simply does not exhibit anything like this sort of *agentive* structure.

* These kinds of DNNs do not involve any kind of recursion or ‘looping’, they just deterministically pass from the activations at the input layer to the activations at the output layer.
* They have no ‘memory’. They cannot change their functioning in light of previous transitions from inputs to outputs. Once the training phase is complete they cannot learn anything or change their input-output functioning in any way in light of new information.
* These DNNs have been trained to (approximately) perform a single desired function.

So there is nothing to be gained, explanatorily speaking, by attributing anything like a goal or desire or an intention to the network.

**AIMS & MESA-OBJECTIVES:**

An important distinction: the ‘aim’ or target used in the training phase vs. the ‘aims’ (if any) that the resulting DNN might have after training is completed.

Comparison with Natural Selection: Evolution operates by blind gradient descent. This process optimizes for reproductive success. But this does not mean that any specific organism produced by natural selection will automatically have the aim/desire for reproducing. Nor does the aim of reproducing exclude the possibility of other competing aims.

Failures of ‘outer-alignment’: when the overall function of the system is not what we wanted. E.g. we wanted it to classify images as either Dogs or Wolves, but actually it is distinguishing images with snow in the background from those that don’t.

Failures of ‘inner-alignment’: when the system acquires intermediate aims/functions in order to perform the desired overall function, but these ‘mesa-objectives’ come to clash/compete with the overall function.

Zerilli seems to assume that such hidden aims cannot arise for systems trained via Machine Learning. He claims that whilst humans may have hidden ulterior motives: ‘ML systems do not have such ulterior motives, of course, so in knowing a system’s beliefs about recidivism it is always safe to assume that its goal will be to give effect to those beliefs.’ (14)

But the possibility of hidden motives (mesa-objectives) is a major topic and concern for AI safety and alignment.

**ARTIFICIAL AGENTS FROM REINFORCEMENT LEARNNG:**

In reinforcement learning the system (‘agent’) can perform various actions within a domain or environment and over the training phase it gradually learns to perform some actions rather than others (in specific situations) via a feedback signal, which either rewards or punishes the agent. Eventually the system learns to (approximately) maximise its reward by taking the (approximately) optimal actions for each situation.

This has been used to train DNNs to successfully play many games – classic Atari computer games such as Pong or Space Invaders

Highly complex modern strategic computer games such as ‘Total War’ and ‘Starcraft II’

AlphaZero can play Go, Chess and Shogi at Superhuman levels.

So if Dennett is right about his 1971-era chess program being usefully interpreted from the intentional stance, then the intentional stance is applicable here also.

**WHAT ABOUT LARGE LANGUAGE MODELS?**

LLMs differ from standard DNNs in various ways.

(a) LLMs have a distinctive ‘transformer’ architecture which alternates layers of specialised ‘attention heads’ with standard ‘perceptron’ layers of artificial neurons.

(b) LLMs *do* involve a kind of looping or recursion – which means that there is effectively a kind of short-term memory of the network’s previous operations.

(c) LLMs deliberately use a stochastic procedure when selecting the next token (word) – known as the ‘temperature’ setting.

(d) LLMs are trained on vast amounts of natural language data, including lots of text that is about human plans, intentions, aims, mental states etc. So the LLM can output natural language sentences that *claim* to express the system’s own desires, plans, intentions etc.

(e) LLMs display a surprising range of abilities to perform tasks that were NOT the (explicit) aim of the training phase – which was simply to predict the most likely next word (token) in a sentence.

“We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4’s performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4’s capabilities, we believe **that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence** (AGI) system.’ (Bubeck et al, 2023 bold-type added)

‘Equipping LLMs with agency and intrinsic motivation is a fascinating and important direction for future work. With this direction of work, **great care would have to be taken on alignment and safety per a system’s abilities to take autonomous actions in the world and to perform autonomous self-improvement via cycles of learning**.’ (ibid p92, bold-type added)

(f) LLMs designed for public consumption as chatbots, such as ChatGPT (as opposed to GPT4) have a second training phase, RLHF, that is not scored according to the accuracy of next-token prediction but rather according to whether Humans find the output ‘helpful and informative’. So ChatGPT is actually the result of two different training-phases with two different optimisation targets.

Is the intentional stance applicable to LLMs?

* LLMs do not exhibit a single stable, coherent set of beliefs & desires over time.
* LLMs do not have any long-term memory outside of the current context window.
* LLMs will sometimes correctly answer a factual question in response to a prompt, but can give a totally different incorrect answer in response to an almost identical prompt with only a tiny, subtle change in phrasing.
* Even within a single conversation (context) LLMs will sometimes forget or contradict what they said in a previous answer.
* LLMs can seem to exhibit wildly different ‘personalities’ according to how they are prompted.

Given that LLMs are disposed to express such a variety of inconsistent claims on almost any topic and seem to exhibit no stable *long-term* aims/desires, it is very doubtful that *the system as a whole* should be interpreted via the intentional stance.

‘GPT-3 does not look much like an agent. It does not seem to have goals or preferences beyond completing text, for example. It is more like a chameleon that can take the shape of many different agents. Or perhaps it is an engine that can be used under the hood to drive many agents. But it is then perhaps these systems that we should assess for agency, consciousness, and so on.’

[David Chalmers, from Daily Nous blogpost, 30/7/2020]

**FRAGILE SIMULACRA:**

Although the LLM considered over multiple contexts/conversations does not seem to display stable rational beliefs, desires or plans, if we just consider the LLM’s output during a single context-window we find that *sometimes* we do get behaviour which seems to express stable and approximately rational beliefs and intentions. I.e. the LLM can produce a short-lived linguistic *simulation* of a rational agent with coherent beliefs and plans and intentions.

People in the AI literature have begun referring to these linguistic simulations of agents as ‘simulacra’.

The LLM does not care what kinds of agents it simulates – indeed it can simulate multiple conflicting agents arguing with each other! In response to the right prompt, GPT will equally happily simulate a crazy conspiracy theorist, or simulate David Chalmers!

[GPT3 producing a convincing imitation of Chalmers: <https://www.facebook.com/howard.wiseman.9/posts/4489589021058960> ]

Analogy: as the Game of Life rules are to the objects (‘gliders’, ‘oscillators’ etc) that can be produced given an initial pattern, so the core LLM is to the simulacra that can be produced given an initial prompt.

GPT4 can be used to play chess. Opinions vary as to how highly GPT4’s chess playing should be rated – somewhere between 1400 to 1800 on the ELO scale.

However, to get it to play chess this well requires very careful prompt engineering. GPT4 can easily prompted into simulating someone who does not know the rules of chess, or who plays illegal moves, etc.

To the extent that GPT4 can be successfully prompted into playing a coherent game of chess, then it seems that this specific simulacrum of a chess player could be predicted and explained from the intentional stance. However, given the fragile and ephemeral nature of these simulations of a reasonable chess player, we should be much more cautious about using or depending on the intentional stance compared with a classically coded chess-playing program.

**MORE STABLE ‘LANGUAGE-AGENTS’:**

SO: without any external memory or APIs etc, the agent-like simulacra that LLMs produce are ephemeral and fragile.

But, as Chalmers’ suggested, things are different when the LLM is a component within a larger system.

If you supplement an LLM with an external memory and with long-term, stable goals, then you can get something that behaves much more clearly like a single approximately rational agent which robustly endures over time and over different contexts/tasks.

See: Park et al (2023) ‘Generative Agents: Interactive Simulacra of Human Behavior’.

This kind of architecture is *not* just being used to control virtual computer game characters. There are already a range of similar architectures that are available for public use: agentGPT, AutoGPT, babyAGI etc. These systems can be used to perform real actions in the real world via the internet. E.g. to invest money, to start a company, to plan an advertising campaign or a political campaign, etc.

Another example is ‘Cicero’ (Bakhtin et al, 2022) which can play the game Diplomacy at the level of a top 10% human player. (Cicero combines a strategic planning unit, trained via RL, with a Language model than can communicate with other players.)

**DANGER/SAFETY:**

Goldstein & Kirk-Giannini have an excellent paper discussing these kinds of ‘language-agents’: “Language Agents Reduce the Risk of Existential Catastrophe” (forthcoming in *AI & Society*)*.*

They argue that since these Language Agents are so readily interpretable and explainable they should be easier to align and so they are a less dangerous form of AI.

I agree that an autonomous agent AI whose plans and memories are explicitly written in natural language is probably safer than a comparably powerful agent AI that is more opaque or inscrutable.

BUT: I think that creating *any* kind of autonomous agentic AI is dangerous and irresponsible.