



Why artificial intelligence needs sociology of knowledge: parts I and II

Harry Collins¹

Received: 16 December 2023 / Accepted: 23 April 2024
© The Author(s) 2024

Abstract

Recent developments in artificial intelligence based on neural nets—deep learning and large language models which together I refer to as NEWAI—have resulted in startling improvements in language handling and the potential to keep up with changing human knowledge by learning from the internet. Nevertheless, examples such as ChatGPT, which is a ‘large language model’, have proved to have no moral compass: they answer queries with fabrications with the same fluency as they provide facts. I try to explain why this is, basing the argument on the sociology of knowledge, particularly social studies of science, notably ‘studies of expertise and experience’ and the ‘fractal model’ of society. Learning from the internet is not the same as socialisation: NEWAI has no primary socialisation such as provides the foundations of human moral understanding. Instead, large language models are retrospectively socialised by human intervention in an attempt to align them with societally accepted ethics. Perhaps, as technology advances, large language models could come to understand speech and recognise objects sufficiently well to acquire the equivalent of primary socialisation. In the meantime, we must be vigilant about who is socialising them and be aware of the danger of their socialising us to align with them rather than vice-versa, an eventuality that would lead to the further erosion of the distinction between the true and the false giving further support to populism and fascism.

Keywords Large language models · ChatGPT · Artificial intelligence · Sociology of knowledge · Studies of expertise and experience · Fractal model of society · Socialisation · Foundations of moral understanding

1 Part I early socialisation, moral understanding and ChatGPT

The field of artificial intelligence (AI), when it is trying to reproduce human intelligence—what I’ll call ‘the scientific problem’—is dominated by mathematicians, neuroscientists and philosophers, thinking about brains.¹ It is not that sociologists are absent but that they tend to look at AI’s reception, interpretation, and impact rather than contributing to the science: few use their understanding of human knowledge to

consider what is needed to mimic human intelligence. And yet, sociologists know quite a bit about human intelligence, quite a bit that AI *scientists*, understood as a subgroup of *knowledge scientists*, could use. One thing they know that is hard for those who focus on the brain to grasp, is that human intelligence is collective—it is property of societies, not individuals. Here I am going to try to show what an understanding of the social could contribute to the understanding of the problems of large language models (LLMs) such as ChatGPT and a little of what that AI project can reveal to those interested in what an understanding of the social can contribute to an understanding of knowledge.²

¹ AI has many faces, most activity going into building useful tools. Here I am concerned with the cutting edge of AI as it approaches nearer to mimicking or reproducing human-like intelligence. Collins 2021, surveys the different claims made of AI and the various responses of critics to these claims.

✉ Harry Collins
CollinsHM@cf.ac.uk

¹ School of Social Sciences, Cardiff University,
Cardiff CF10 3WT, UK

² I say refer to ‘those interested in an understanding of the social’ rather than ‘sociologists’ since many members of the sociological profession are more concerned with social inequalities or injustices than the idea of the social as a basic constituent of the world. At the same time, a dominant trend in science and technology studies, led by the late Bruno Latour, takes its lead from literary theory and the associationism of Gabriele Tarde rather than Durkheim’s ‘treat social facts as things’. Of course, there are others, such as philosophers like the later Wittgenstein (1953), Bloor (1976, 1983), whose interest in the nature of knowledge leads to a focus on the nature of the social.

These recent developments in AI, I will argue, have brought a new emphasis to the fact that human intelligence depends on moral intelligence. Human intelligence is collective and you cannot have collective intelligence in the absence of moral integrity because without moral integrity there cannot be productive social interaction.³ Drawing on a schematic understanding of society, I will try to explain LLMs' tendency to provide anti-social responses and also fabricated responses—what their creators refer to somewhat misleadingly as 'hallucinations'. There are many bizarre stories in circulation about hallucinations and since ChatGPT is free to access one can easily confirm the tendency for oneself. For example, I asked ChatGPT to tell me who I am and it provided a convincing looking answer including a list of six books I had written. But one of these had a mistake in the title and an incorrect publication, date, one didn't exist as far as I could see, and two were the modified titles of books written by other authors. All of this was presented in a convincing format and writing style which would not cause the innocent reader looking for information to question it. To slip into anthropomorphic language, ChatGPT simply does not 'know' the difference between the true and the false—it does not 'know' it is fabricating, and it does not 'realise' that there is anything bad going on. If it was human, we might say it was a psychopath in so far as this term intimates lack of empathy for others and lack of remorse for anti-social actions. But it is worse: it does not know what truth is. I think one can explain why ChatGPT is like this by thinking in terms of human socialisation.

'Socialness', it has been suggested, is a basic characteristic of humans in the sense that 'consciousness' is such a basic characteristic. In 1998 this author defined socialness and defined 'applied meta-sociology', as follows:

"Socialness is the capacity to attain social fluency in one or more cultures." If one has social fluency, one has social capabilities and one can follow rules in the Wittgensteinian sense. ... Applied meta-sociology examines what entities who possess socialness can do and what entities without socialness cannot do. (Collins 1998, p 497)

Applied meta-sociology allowed us to see where the artificial intelligence of the time was failing: it was replacing the ability to become socialised into social groups with mechanical rule-following. The rules encoded in AI programs had been extracted from humans, either from the programmer's reflecting on their own understanding of the world or, in the case of the expert systems boom of the 1980s and '90 s, by

³ Shapin's 1994, *Social History of Truth* makes this point in respect of studies of science. We can look at large language models as illustrating what could happen if we remove the moral dimension from knowledge creation.

programmers interrogating other humans who were experts in various esoteric domains. In both cases it was a matter of trying to describe the social from an external perspective rather than acting according to what was known from the inside—the actions being guided by 'actors' categories'. Machines programmed this way fail Turing Tests designed to expose the tacit knowledge embedded in actors' categories, the ability to understand legitimate rule-breaking, and so forth.⁴ But recent developments in AI have shifted the focus of the sociological critique because the machines are now undergoing something much closer to human socialisation: they are embedding themselves into human language in a much more human-like way. The refocussed critique deployed here is simpler and more schematic than the earlier approach. Here, I am going describe the process of socialisation in a simplified way, capturing how members of society attain fluency in language and describing how this differs from the way ChatGPT and similar 'large language models' (LLMs) attain fluency. I am going to use that schematic difference to help us understand both ourselves and large language models.

1.1 The fractal model of society and human socialisation

The 'fractal model of society' is central to a research program known as Studies of Expertise and Experience (SEE). The ideas have been building in the social studies of science domain since the 1970s, developing particularly strongly since the turn of the century.⁵ In one way, SEE has to be seen as a friend of the latest developments of AI because SEE

⁴ At this juncture this author was developing critiques of AI that turned on its incorrect concept of human rule following, well exhibited in continual failures to pass Turing Tests. Human rule-following often involves rule-breaking (Collins 1990) and human rule-following often requires instantiation of the same actions with different behaviours (polimorphic actions – Collins and Kusch 1998). In 2012, Levesque, Davis and Morgenstern reported on a Turing Test conducted in New York City to which all the most advanced artificial intelligence devices were invited. The test was based on Winograd Schemas, which are designed to require commonsense knowledge if correct responses are to be forthcoming. AI companies either refused to take part or their machines performed abjectly (see also Collins 2018a).

⁵ The more recent developments in social studies of science described here turn on a shift from the social construction of truth to the analysis of expertise signalled by a paper published in 2002 by Collins and Evans. The programme has been carried forward in many publications (eg books by Collins and Evans 2007; Collins 2017; Caudill et al 2019; Collins et al 2022a, b) and involves an international group held together by a regular, hybrid seminar held at Cardiff university, which meets in person for an international workshop when possible (at least annually before Covid). This organisational focus means, among other things, that there are 'pathological' number of self-citations in this piece, exacerbated, in the case of the large number of joint works, by the contingencies of alphabetical priority. All mistakes and infelicities remain my responsibility.

stresses the vital importance of the linguistic component of socialisation—referred to as *interactional expertise*—which is to a large extent, a foundation for, and can be a replacement for, the practical component. SEE argues that in some senses, ‘language contains practice’. Thus sufficiently thorough immersion in a local language can produce a level of socialisation that is indistinguishable from socialisation as a whole, certainly when tested by ‘Imitation Games’, which are Turing Tests with humans from one social group trying to mimic members of another social group.⁶

Figure 1 represents a society indicating that humans come to know what they know via socialization into a variety of social groups. Of course, the groups shown in the figure are a small selection of the indefinite number of groups that can be found in a real society. The model is inspired by the metaphor of the fractal, with some technical features in common with fractals as mathematicians think about them. The method of socialization into groups and the method of testing for socialness of groups are the same at every level from top to bottom and this is fractal-like. The fractal is also more literal than metaphorical when it comes to the way smaller groups lower down in ‘the fractal model of society’ are embedded in the upper groups: they are both embedded in them but still constitute them even though they are separately identifiable.⁷ The notion of ‘socialness’, mentioned earlier, implies that socialization depends on the acquisition of tacit knowledge and cannot be straightforwardly replaced by bodies of information.

Cauliflowers, like societies, are examples of physically instantiated fractals. A cauliflower has florets within florets within florets embedded in a cascade yet without the florets there is no cauliflower. Still, you can identify the sub-florets down to any level. As with the cauliflower, a society is constituted by many sub-societies—which we’ll call ‘groups’—each embedded within each other but each characterized by their own ever more specialized ways of being in the world. At the same time there is no society as a whole without the sub-societies continually interacting with and continually reconstituting the entire organism. Some ways in which the technical idea of the fractal is *not* exactly applicable to society include that human individuals can belong to many groups at the same level and the mutual embedding is multi-dimensional.⁸ In physically instantiated fractals, as opposed

to mathematical abstraction, the fractal-likeness ends at the bottom as a few individual cells in the case of a cauliflower and a few individual humans in the case of societies.

Becoming an individual member of such a society is a matter of socialization into a sub-set of the society’s groups—scientists, chemists, cricketers, protestants, stamp collectors, and so on, but always including the top level. The top level is the location of the typical *ubiquitous expertise* which characterize that society in particular, such as fluency in the native language, understanding the moral code, what counts as the difference between clean and dirty in that society, and, in ‘Western societies’ a basic understanding of political choice. Here we will concentrate on language and moral code, which we can call ‘basic ubiquitous expertise’.

Somewhat arbitrarily, we’ll divide socialization into three stages (see Table 1). Basic ubiquitous expertises, are mostly acquired by the newborn and the toddler during what we will call the ‘primary socialisation’ that takes place in the family.⁹ Secondary socialization takes place when the child leaves home and starts to learn more specialist skills, including more about clean and dirty and other social skills, and written language which is learned at primary and secondary school. The child will begin to choose, or be directed into, a small subset of the more specialist groups below the top ubiquitous level, with membership of the subset beginning to define the person as an individual. Tertiary socialization is a more specialized version which happens in higher education and specialized occupations.

Seen this way, the person is a molecule made up of larger ‘atoms’—the social groups in which individuals are immersed and become socialised. When a thermometer is dipped into a liquid the reading depends on the temperature of the liquid and in the same way an individual’s, say, fluency in a natural language depends on the society in which they are immersed, and the same goes for the many more specialized understandings we acquire through socialization

⁶ See for example Collins 2004; 2011; Collins et al 2006

⁷ The fractal model can also be seen, with a bit of a stretch, as a generalisation of Durkheim’s way of seeing the relationship between the professions and society. Durkheim developed this idea in a series of lectures starting in the 1890s and published in various editions as a book, entitled *Professional Ethics and Civic Morals*, the latest being in 2019. See also Collins 2023.

⁸ Not all sets of people with similar characteristics are a ‘group’. A group is characterized by the tacit knowledge that must be acquired in

Footnote 8 (Continued)

becoming a member. So, people who have brown hair are not a group and neither are people who wear shoelaces: there is no tacit understanding associated with those characteristics (at least, not in the UK and US). The difference between groups and sets can be understood by thinking about the Turing Test, or its immediate predecessor, the imitation game which involved a hidden man pretending to be a hidden woman while both were interrogated by a judge. It should be easy to pretend to be a member of a ‘set’ simply by lying but pretending to be a member of a ‘group’ when interrogated by a member of that group is much harder because of the tacit understandings involved. Work on a methodologically refined version of the Imitation Game was supported from 2011–2016 by European Research Council Advanced Grant (269,463 IMGAME) €2,260,083 ‘A new method for cross-cultural and cross-temporal comparison of societies’.

⁹ Ubiquitous expertises such as an understanding of democracy may be acquired quite a lot later in the process of socialisation.

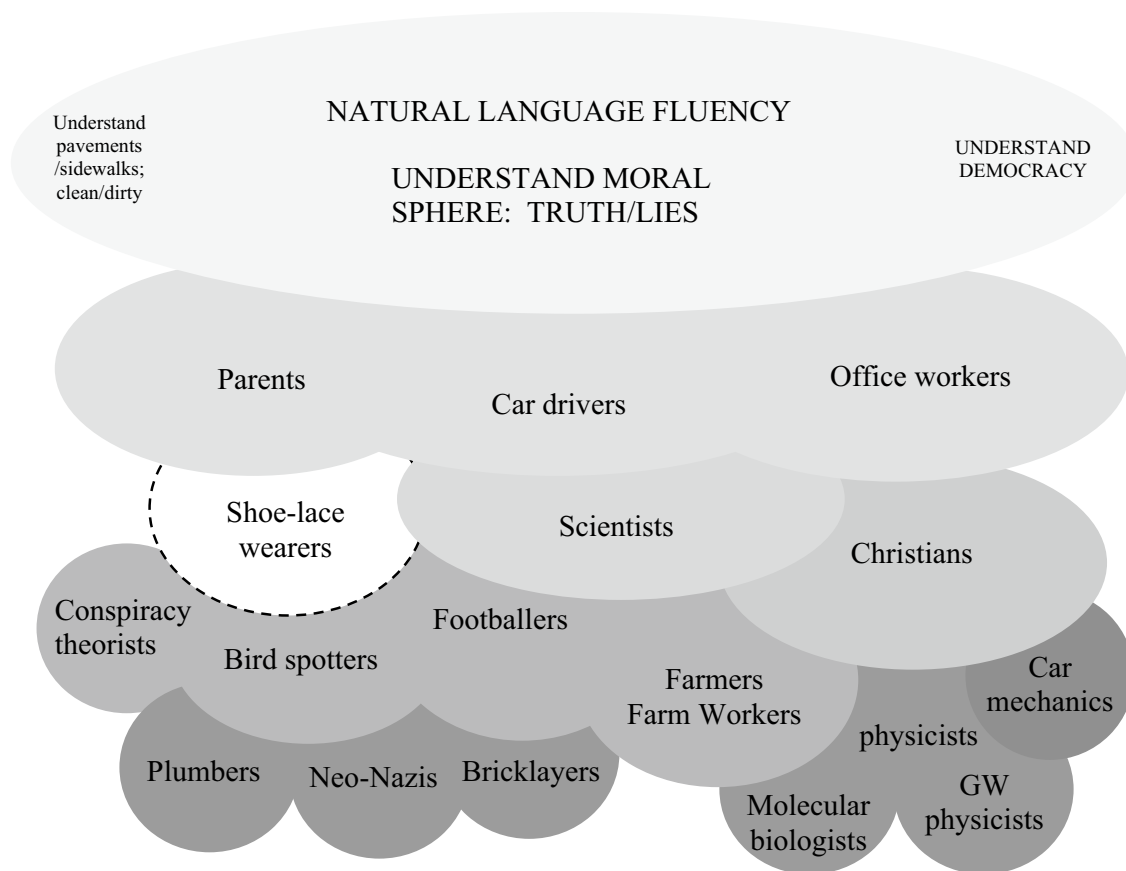


Fig. 1 Some groups in the fractal model of society (eg UK or US). (This figure has been reproduced with minor variations in a number of publications.)

into groups.¹⁰ AI is beset by the idea that the brain does the work of making knowledge whereas most of what the brain does is extract knowledge from the groups in which it is embedded. This is another way of expressing the basic insight of the sociology of knowledge.

1.2 GOFAI, NEWAI, and the ‘socialisation’ of large language models such as ChatGPT

The dominant mode of artificial intelligence work up until the last ten years or so was what was known disparagingly by its critics as ‘good old-fashioned AI’ or GOFAI. It involved humans writing programs meant to reproduce human thinking. The expert systems boom of the 1980s and 90s was typical of this approach: AI researchers would interview human experts and try to extract the knowledge from their





¹⁰ SEE diverges from psychological and philosophical analyses of expertise by treating all expertises as matters of socialisation into expert communities. Expert communities have their own specialist languages, though the fractal metaphor encourages these to be seen as a local ubiquitous expertise when viewed from within the specialist group. See, eg, Collins 2018b

heads and reproduce it in computer code. The term ‘GOFAI’ was invented by philosopher John Haugeland, who was a student of Heideggerian philosopher and well-known critic of AI, Hubert Dreyfus. Dreyfus became notorious for predicting, wrongly, that no computer would ever beat a grand master at chess.¹¹

Dreyfus argued that grandmasters viewed the chess board holistically in some not fully explicable way, whereas GOFAI could only calculate relative advantage based on an unfolding tree of possible moves and counter moves. Such a tree explodes exponentially, soon exhausting human capacity to calculate ahead and exhausting computers’ capacity only a little further down the line. It was thought computers were certain to win using this method only if they could calculate the outcome by following the expanding move-tree all the way to the end of the game, as in tick-tack-toe (noughts and crosses). But in chess this would require a computer many times bigger than the universe. Computers were getting faster and chips getting cheaper, and it turned out that

¹¹ See for example Dreyfus’s well-known 1967 book and the 1992 reprint.

Table 1 Three stages of human socialisation

Stage of socialisation		Where and What?	
		Location	What is acquired
Primary		Family home	Native spoken language
		Local tribe	Moral concepts
Secondary	 	Primary then high schools	Written language Early technical concepts
Tertiary/professional		Higher education/apprenticeship	Specialist language and skills

only a few years later they would be able to calculate a little further down the line still, and, combining this with ways of estimating the strength of the position on the board at any one time, it would be enough to beat any human even without getting anywhere near the end of the game. The fact that you had to go only a little further forward with mechanical calculation to win nearly every time, took everyone by surprise and that is why Dreyfus was not as wrong as it appears since he was basing his arguments on what was known at the time about human and computer chess players. Since then, however, all this has been overwhelmed by 'NEWAI'. NEWAI—'new artificial intelligence'—is my term for computers based on neural nets and comprising deep learning together with large language models, both of which have been hugely successful in recent years. NEWAI in its deep learning form, is the undisputed champion at games with exhaustively defined rules, such as chess. In such cases a deep learning machine can acquire, in a day or two, the ability to beat anyone by playing millions of games against itself and noting the moves that eventually lead to victory. This is a bit more like the human way of acquiring expertise at chess and such like, though it massively exceeds human abilities in terms of speed and volume.

Deep learning and large language models (LLMs) are, as intimated, both based on neural nets. Neural nets are an old idea, going back to the 1950s and 60s, which were starved of attention by champions of GOFAI such as Marvin Minsky, who persuaded the AI community that they were a dead end. Their startling revival a few years into the 21st Century started with deep learning and a huge leap forward in competent language translation and the like. This was supported by massive increases in relatively cheap computer power, possibly driven by the games industry. From a sociological point of view, one reason why NEWAI holds far more promise than GOFAI is that it is capable of teaching itself

new rules rather than needing to have a human insert all the instructions so it can learn from a changing society. This invites a comparison with human socialisation,

Neural nets are said to be modelled on the human brain. Neural nets work by making random guesses at what they should be outputting when they are given some input such as a matrix of pixels (say a matrix that humans would read as the letter 'A'), and having those guesses confirmed or rejected by some form of supervision. Every now and again one of those guesses will be better than the other guesses and then the net has to be 'told' that it is on the right lines. It will change its internal state by adjusting the weights that ease passage along the pathways between its artificial neurons so as to reinforce the likelihood that in the future the more successful guess will be more likely to be made than the less successful guesses. This process is repeated over and over, improving the pattern recognition abilities of the neural net with every iteration, until good performance is reached. As can be seen, reinforcement of better guesses—supervision—is vital. One form of supervision is quite deliberately designed and controlled by humans who know where they want the recognition abilities of the machine to go. On the other hand, there is so-called, 'unsupervised learning', where the net is presented with data and is given, say, a minimal, fixed instruction to separate what it 'sees' into discrete entities which recur in repeatable ways. It is said that even with no supervision beyond this, neural nets presented with a jumble of badly handwritten numbers can eventually separate them into the correct ten classes. Some computer scientists think that this is the crucial form of learning and that it reflects what humans (and other living creatures to a lesser extent), did in the course of evolution as they came to be able to understand and manipulate the material world around them. But there is an in-between kind of supervision which involves the machines, by themselves,

finding some repeatable order in the world around them but an order that has been invented by human cultures. This is a kind of human supervision but one that is immanent in the cultural artifacts which are presented to the computer for analysis. An obvious example is a neural net learning to read a printed or digitised language like English, graduating from the letters and numbers to whole words and eventually sentences. The net is finding a pattern in the world, but it is a pattern that has been put there by humans. I call this ‘implicit supervision.’¹² Going back to the recognition of handwritten numbers, one can see this as implicit supervision rather than no supervision though it’s no supervision when the objects being separated are found in nature as with the proposed mechanism of evolution. On implicit, or not so implicit supervision, Alan Blackwell has expressed the point graphically when describing the way deep learning comes to identify pictures of objects:

. . . thousands of people are paid pennies to create a ‘ground truth’ by providing labels for large data sets of training examples. . . . In this case, the ‘objective function’ is no more or less than a comparison of the trained model to previous answers given by the [humans]. If the artificially intelligent computer appears to have duplicated human performance, in the terms anticipated by the celebrated Turing Test, the reason for this achievement is quite plain—the performance appears human because it is human! . . . The artificial intelligence industry is a subjectivity factory, appropriating human judgments, replaying them through machines, and then claiming epistemological authority by calling it logically ‘objective’. (Blackwell 2015)

Large language models, the other kind of NEWAI, depend on neural nets but for the purpose of this analysis they can be thought of as ingenious extensions of the predictive text found on a word processor or mobile phone. Predictive text offers a word or two as potential continuations of what you are writing but LLMs continue the process to paragraph length or longer.

Thus, I now type into my computer: ‘what is the highest mountain in the world?’ As I reach the ‘t’ in the penultimate word my computer completes the question for me. The ‘intelligence’ in the word-processing program I am using ‘knows’ enough about the English language to predict that having got that far in the sentence there was a very good chance that a typical English-writer would complete the sentence in that way, so it offered that continuation to me to save me the trouble of typing it myself. It is all a matter

of statistical analysis of corpuses of written text and assembling a list of probable continuations. ‘he world’ follows the last ‘t’ in ‘What is the highest mountain in t . . .’ on a very large number of occasions in the corpus whereas, say, ‘yplop grubston’ probably never follows it. Nobody would write ‘What is the highest mountain in typlp grubston?’, however large the corpus of written text, and even if they did (it has just appeared once in some written text!), it would hardly ever show up, so the way things go, the statistics encourage the predicted completion of the sentence to be ‘he world’.¹³

There are lots of possible variations in the way LLMs work. For example, LLMs don’t always choose the *most* probable continuation for their prediction. This, it seems, would produce text that is not very lively, so slightly lower probability continuations are chosen which give rise to more interesting writing and allow different answers when the program is asked the same question repeatedly. Then there is the matter of how many words are taken into account from the corpus when calculating the probabilities. The program can look at longer or shorter strings of words taken from the corpus both before and after the next word to be predicted. The programs also use something called ‘the transformer’ which analyses the relationship of words in the actual text being written in a more complex way that pulls out the focus of an inquiry (see Madhumita 2023). The sociological point being pursued here does not depend on these details, however but, as intimated, takes a more schematic perspective on the notion of socialisation.¹⁴ What is important for the sociological analysis is what sources

¹³ I don’t know what corpus of text my computer has analysed in order to come up with that continuation. It could be continually analysing some vast corpus in real time via the internet but I don’t think my internet connection is fast enough to enable that. I am guessing that the word processor is pre-loaded with the statistical probabilities harvested from some earlier analysis of a vast corpus. But it turns out that my machine is also doing some real-time analysis on the little corpus being newly built right here and now, maybe just specific to this document: I write some new prose, thus: typlp grubston is a beautiful country, a magnet for tourists who come to climb typlp grubston’s impressive mountains. Some of the highest mountains anywhere are found in typlp grubston. People often ask, What is the highest mountain in typlp grubston?

That time I got to the ‘g’ in the last word and the predictive text offered the completion ‘...rubston’, so the probabilities within this laptop’s program are being modified by my real-time writing propensities. I doubt if the influence of ‘typlp grubston’ is getting beyond my computer but if this text I am writing is published and comes to be included in the huge corpuses upon which other predictive text probabilities are based, ‘typlp grubston’ might have a slight influence on everyone’s predictive text.

¹⁴ More technical treatments intended for a wider audience can be found at Madhumita (2023) (<https://ig.ft.com/generative-ai/>) and Guinness 2023 (<https://zapier.com/blog/how-does-chatgpt-work/>), with a more technically demanding treatment at Wolfram 2023 (Wolfram, Stephen (Feb 14, 2023). "What Is ChatGPT Doing . . . and Why Does It Work?").

¹² For a more complete explanation of these kinds of supervision in the context of a discussion with pioneer, Geoffrey Hinton, see Chapter 6 of Collins’s 2018a, *Artificial Intelligence*.

of information ChatGPT uses. Guinness 2023, provides a description of what ChatGPT draws on:

All the tokens [words or parts of words] came from a massive corpus of data written by humans. That includes books, articles, and other documents across all different topics, styles, and genres—and an unbelievable amount of content scraped from the open internet. Basically, it was allowed to crunch through the sum total of human knowledge to develop the network it uses to generate text (<https://zapier.com/blog/how-does-chatgpt-work/>)

This quotation is informative in two ways: it indicates the huge processing power of contemporary computers and how they can be used to handle almost inconceivably large bodies of data, but it also exhibits a revealing but common mistake, confusing what is found on the internet with the sum total of human knowledge, a point to which we will return.

One other feature of NEWAI, reaffirmed by the quotation, which gives it a huge advantage over GOFAI, or, at least, so it would appear at first sight, is that NEWAI teaches itself, at least, up to a point. With GOFAI, humans extract the rules from human activity as far as they can and insert them into programs. This means that the programs are frozen in time whereas human language (for example) is continually changing as society changes so a frozen rule-base is soon out of date unless it is continually updated by humans. NEWAI has the capacity to keep up with a changing society, in so far as changes on the internet reflect changes in society. This potentially gives NEWAI a big advantage if one is concerned with what I am calling the scientific problem and this is another of the features of NEWAI that invites comparison with human socialisation.

But, as it turns out, ChatGPT (and its successor, GPT4), are ‘pre-trained’ (GPT stands for ‘Generative Pre-training Transformer’) with the training having ceased in 2021. The programs not only cut off access to the internet at 2021 but also have no memory of their interactions with users after 2021 once any particular interchange has come to an end, so, the major advantage described above has been discarded! This is strange but is almost certainly revealing in ways that will be suggested.

1.3 Human knowledge: research science and primary socialisation

We now look at human socialisation at the extreme ends of the fractal, babyhood and specialist research science. Going back to the acquisition and establishment of human knowledge, primary socialisation and that part of tertiary socialisation that comprises research science have three things in common: (a) they are both about learning about or establishing the existence of things that are new to the learner; (b) in

both cases this is accomplished in small, bounded, trusting groups relying on face-to-face communication—the family in primary socialisation and the core-set or core-group in research science; and (c) in both cases they depend on moral intelligence and moral integrity—acquired in primary socialisation, all being well, and additionally reinforced in the institution of science.¹⁵

The situation in research science is well-understood as a result of historical analysis and field studies of the kind which began in the 1970s, probably triggered by Thomas Kuhn’s, *Structure of Scientific Revolutions*, which was published in 1962. What Kuhn showed was that science was not simply a mechanistic set of procedures inspired by genius but involved cultural variations or ‘paradigms’, the overall sets of taken-for-granted assumptions within which theorisation and experimentation took place. Another vital concept set out by Kuhn still earlier, in 1959, was what he called ‘the essential tension’. This is the tension between radical innovation on the one hand and acceptance of the constraints of working within a paradigm on the other. Both were necessary in science but most of the time it is ‘normal science’—working within a paradigm—that keeps science going, even when it is frontier research that is going on. This was long before we needed to worry about uncontrolled interventions from the internet but, even then, given the constant criticism from the fringes of science or occasionally from inside other paradigms, science can’t develop new concepts and understandings unless there is consensus and trust within the research team, so that everyone in it is working from the same set of assumptions. As it happens the groups which develop new ideas are generally small, and this means they can frequently meet face-to-face, develop and share the new language that embeds the new concepts and procedures, learn to trust each other, readily coordinate a division of labour between practical specialists, and so on.¹⁶ The groups that develop these new ideas are like families or small tribes, knowing each other well, sharing a language and a set of understandings, and careful about admitting strangers.¹⁷

We have been led from describing teams of research scientists to families and small tribes. This is not a coincidence. One cannot have a describable world without enduring objects and concepts. If the descriptions of objects or concepts are continually changing then there are no objects and concepts. I am going to speculate that the stability of

¹⁵ For a discussion of the importance of small, trusting, face-to-face groups in research science see Collins et al 2022a, b

¹⁶ My 45-year-long study of gravitational wave detection shows how all this works. Collins et al 2017 is a field study of the fringe.

¹⁷ The internet is threatening to change all this, reinforced more recently still by the restrictions on travel encouraged by global warming and the pandemic. For the potential effect on science of the restrictions coming in the wake of the pandemic, see Collins et al 2022a, b

the world of things is the first thing a newborn learns in early baby talk with parent or carer consistently repeating the descriptions of simple objects. I am going to speculate that this is how the child learns the basic concept of truth—this object really is called ‘that’ and only ‘that’, every time it is named. Imagine a parent or carer describing nursery objects differently every time the child encountered them, sometimes as blue, sometimes as red, sometimes as ‘rattle’, sometimes as ‘dummy’, sometimes as ‘mother’, sometimes as ‘father’, sometimes as ‘dog’. The child would never learn to name things, would never learn language and, perhaps, would never learn there is stable world or learn the application of truth. In the normal way, this sense of stability and the sense of truth is fostered from the outset and reinforced as the child later encounters the extended family and, perhaps, the local tribe, all of whom see and describe the world in the same way.

In terms of the fractal model of modern societies, families are themselves small, specialised, groups, responsible for socialising baby newcomers into an understanding of ubiquitous expertises. I am going to suggest that the sense of truth and stability comprises the basic building block of moral sense: learning language is learning to tell the truth and telling the truth is the first step in building a moral compass. This is the first step in primary socialisation. Both newborns and research scientists are soon socialised into the sense that their unfolding worlds rest on truth and stability; parents are always telling their children that ‘this’ really is one of ‘these’ while scientists are always engaged in discovering and agreeing about what they are looking at really is.

In the case of families, the proper socialisation comes naturally. It also gives rise to stable sense of society based on the other ubiquitous expertises which are learned as the child grows, such as the sense of clean and dirty and proper behaviour in the society in question, an understanding of democracy, at least within Western democracies, and so on. We know there are cases where a newborn isn’t given the normal care, this usually being described in terms of a lack of love (as in Bowlby’s 1953, book). We also know there will be families which imbue values different to the ubiquitous expertises that form the society in question, but these generally remain outliers, disconnected from society as a whole and, in relatively stable societies, unreinforced by interaction with other family groups or the wider society.

The relationship between core and outlier is different on the internet. On the internet, there is no reason why what count as outliers in society should be unrewarded—indeed the creators of outlying positions set out to attract followers and often become wealthy as a result; the social outliers are often the most active in the digital world. For an LLM, ‘socialisation’ or its surrogate, starts around the wavy arrow shown in column 1 of Table 1: that is, it begins most of the way through the secondary

socialisation of humans, and is restricted to the internet; it is not going to implant the basic building blocks of the moral and is not going to distinguish between values that are formative of that society and values that are outliers: they all look the same on the internet, perhaps differentiated statistically but with no guarantee that the statistics will correspond with the social differentiation found in the society.

In the case of core-groups of scientists, the very existence of the institution of science turns on a collective search for the truth about the observable world—what I’ll call ‘correspondence truth’—and to members of core-groups it soon becomes clear that the collective project will fail unless they cleave to what I’ll call ‘moral truth’, which is an internal state—a determination to do everything possible to tell the truth to fellow groups members others about the substance of their individual observations.¹⁸

The equivalent of the little ‘families’ that are the core-groups at the research frontier of science are similar to ordinary families in that they overlap enough at the edges to give rise to a common morality—the set of norms and values of science—which characterises the sub-group of science as a whole found higher up the fractal than any specialist groups of scientists, but below the ubiquitous expertises. Scientists, like every citizen, draw on the ubiquitous expertises including the idea of truth and the native language, but they have acquired these at an early stage of their lives. The core-groups are like the ordinary families in that they overlap enough to give rise to science’s ‘specialist’ ubiquitous expertises. Once more, there will be outliers. For example, aspects of economics’ mainstream, such as market fundamentalism, may be an outlier science in respect of the norms of science as a whole, being more closely integrated with big business and a certain strain of politics—institutions which do not cleave to truth as a formative value—than with the institution of science.¹⁹

1.4 Large language models revisited

As we have seen, a feature of the communities of research scientists is that they restrict entry so as to defend the stability of the observational worlds they are constructing. In the same way, the family and the small tribe severely restricts the envelope of conceptual opportunities so as to allow the child to experience a stable world. In the case of the newborn, what this also does, we are arguing, is create the basic building block of a moral compass—truth. Apart from some outliers, it also builds the formative values of a societies.

¹⁸ For a recent philosophical analysis see Collins 2023.

¹⁹ For an analysis of the way politics and business influenced ‘neo-liberal’ economics, see Oreskes and Conway 2023.

Thinking purely in terms of socialisation, we can begin to get a sense of the essential difference with large language models. If left to themselves, large language models would read everything on the internet but on the internet there is no stability, no constancy and no truth; rather there is every possible option and opinion. An LLM reads everything and then, instead of being restricted to the trustworthy, it calculates probabilities based on all the opinions it finds. This works brilliantly when it comes to writing fluent English (or some other native language), because nearly all the contributors of views it is examining are fluent writers. It can even extract, using statistical tools, a selection representing one native language or another and translate between them, and it can select certain styles of writing located within one language and associate them with an author. But it fails when it comes to the substance because there is no well-organised substance.

Even within human science, the hard job is to know how to pick the few papers that need to be studied from the snow-storm of publications and preprints that might initially seem worth of consideration. Then we move outwards to works that emerge from the fringe and alternative paradigms. In academia in general, a tool like Google Scholar when tasked with identifying even a single author's publications, finds all kinds of strange things. For instance, I think I have a publication list about 250 strong whereas Google Scholar has 503 items accredited to me: of the extra 250 or so, about 40 have been cited only once and the rest not at all. Most of these c250 I do not recognise. After this, to get 'knowledge' as opposed to text from the internet, we have to know how to reject conspiracy theories, click-bait and organised misinformation and disinformation.

It will be argued by some that to criticise LLMs because of their deficient socialisation is to miss the point. Isn't it the case that ChatGPT and other LLM's are not conscious and don't understand meanings, being nothing other than statistical engines, and that explains their problems? There is a huge philosophical debate about the importance of conscious to AI and to humans, but let it be the case, for the sake of argument, that conscious understanding is important and let us see if the argument from socialisation still stands.²⁰ Allow a newborn human, miraculously born with the ability to read, to be given the same introduction to the world as an LLM: access to the entire internet starting a good way through secondary socialisation as represented by the wavy arrow in Table 1. Neither LLM nor newborn would have any primary socialisation. Instead, they would, at best, absorb the secondary and tertiary socialisation of groups represented on the internet. In the case of the miraculous, but still relatively limited in power, newborn, we have no

idea how it would choose what it would read. In the case of the LLM, it would read the output of every group in the fractal model, not only in the UK and US but of everyone in every society that makes use of the internet, for good or ill. Science and primary socialisation work because they restrict the available perspective but a human who started life like an LLM, with no guidance as to what to read from the huge amount available to it would also form no stable world. The argument from socialisation is just as valid even if we insert human-like consciousness but align the process of socialisation with that of LLMs! The word 'align' is going to continue to be important.

1.5 The return of GOFAI

The creators of LLMs have discovered these problems for themselves. It turns out that LLMs cannot simply be let loose on the world to socialise themselves without disaster.

Guinness (2023) points out:

Of course, GPT's initial neural network was entirely unsuitable for public release. It was trained on the open internet with almost no guidance, after all. So, to further refine ChatGPT's ability to respond to a variety of different prompts in a safe, sensible, and coherent way, it was optimized for dialogue with a technique called reinforcement learning with human feedback (RLHF). Essentially, OpenAI created some demonstration data that showed the neural network how it should respond in typical situations. From that, they created a reward model with comparison data (where two or more model responses were ranked by AI trainers), so the AI could learn which was the best response in any given situation. While not pure supervised learning, RLHF allows networks like GPT to be fine-tuned effectively. (<https://zapier.com/blog/how-does-chatgpt-work/>)

OpenAI, the team that developed ChatGPT are quite frank about their interventions into the machines' abilities. Here are the creators of GPT-4 (a more advanced version of ChatGPT but one that requires a subscription) writing in their technical report (OpenAI 2023) of their aim, which, as can be seen, is to make the machine align with *a wide swath of users' values*. The programmers are acting on behalf of society as surrogate parents and reintroducing aspects of a primary socialisation retrospectively; they are providing the machine with a hand-crafted, surrogate moral compass:

GPT-4 has various biases in its outputs that we have taken efforts to correct but which will take some time to fully characterize and manage. We aim to make GPT-4 and other systems we build have reasonable default behaviors that reflect a *wide swath of users' values*, allow those systems to be customized within

²⁰ For arguments about the place of consciousness in human and other life see, for example, Chalmers 1996 and Reber 2018.

some broad bounds, and *get public input on what those bounds should be* (p11, my stress).

Here is an example that can be found in a technical report which shows how a response about bomb-building was disallowed ‘by hand’ (Table 2).

Here are a couple more examples of what I presume is retrospective socialisation, which emerged from my own interactions with ChatGPT (Table 3):

But when asked if it had a moral compass, ChatGPT ‘denied’ it in spite of its claim to adhere to ethical and moral principles, seemingly resolving the huge philosophical debate about the significance of conscious understanding along the way (Table 4):

The mechanism used to produce the ethical answers seems, according to a story in *TIME Magazine*, (Perrigo 2023), to hark back to the description mentioned earlier (Blackwell 2015), of the labelling of examples by ill-paid human readers located in distant countries. In this case the workers were paid to label examples of toxic text. According to Perrigo, ‘To get those labels, OpenAI sent tens of thousands of snippets of text to an outsourcing firm in Kenya’. The newly labelled examples were then used to train the program in how to react when such things are encountered. So, this, once more, appears to be certain human judgements, common to Kenya and ‘Western countries and therefore likely to be fairly high in the fractal model and acquired in fairly early socialisation, being inserted by humans into the responses of the LLMs as retrospective socialisation.’²¹

1.6 Could LLMs acquire a moral compass?

What seems to follow is that if an AI it is to reproduce a humanlike moral compass, and if it is automatically to make moral judgements that align with societal values in the same way as a typical family, it will need a humanlike primary socialisation. It will need to be ‘brought up’ in a regular family rather than have a set of judgments inserted retrospectively by human adults.²² The approach to implanting human-like socialisation into computers, if they are to mimic humans and make a contribution to the scientific problem, will have to be like that which was once used to try to teach humanlike language to apes, such as the famous ‘Washoe’ and ‘Nim Chimpsky’. This would need some advance in neural net technology as the machines would have to understand speech not just text, and recognise objects at least as well

as a physically challenged child, but these competences no longer seem beyond the reach of future technology. However technically proficient the machines become in doing this, it remains that, counter-commonsensibly, a better socialisation is a narrowly restricted socialisation. As it happens, neither Washoe nor Nim Chimpsky got very far in language learning (nor acquiring a moral compass), almost certainly because they didn’t have the right kind of brain and speech related anatomy (as with domestic cats and dogs which are offered all the advantages of human socialisation).²³ In this respect computers are more promising. Human-like primary socialisation of computers might not lead to success in moral development and value choice either, but it is hard to see how any other approach could work. As regards AI scientists, it is being argued that an understanding of the first principles of human intelligence must start with an understanding of the social.

1.7 Conclusions and consequences

The model of human intelligence that informs much artificial intelligence work is too much based in the individual brain and an old-fashioned view of science: ‘the brain is the repository of intelligence, which gives rise to scientific knowledge, and the more of it the better’. This model leads to fear of the singularity: ever more powerful computers with ever more scientific knowledge, will construct still cleverer computers, which will eventually so exceed human intelligence that we will be lucky if they keep us as pets. The potential is there to see when we compare the capabilities of human brains with the capacity of computers when it comes to, for example, professional examinations. It is no surprise that computers outstrip us, and will soon vastly outstrip us, at legal bar examinations, where they can read and analyse every legal word encapsulating every legal precedent, and that they threaten to surpass us in some branches of science, such as molecular biology or chemistry, where what is needed is the ability to handle and manipulate huge numbers of variations of well-understood and uniformly accepted procedures.

But most human knowledge is not like this. Most human ‘knowledge’ is not knowledge at all. Most of human knowledge, if knowledge is thought of as what is found in printed or digitised sources, is useless, wrong, conflicting, confusing, deliberately designed to misinform, or to bring about certain ends that most of us would find undesirable. The more we rely on computers to provide our ‘knowledge’, the more will we be vulnerable to this dangerous penumbra.

²¹ It is interesting that basic moral categories can be safely assumed to be common to many societies.

²² This is not to say that some human socialisation does not take the form of correcting ethical and other kinds of mistake and misjudgement: humans too are always learning from wiser or more experienced persons, but they already have the basic building blocks of truth and consistency to support these corrections.

²³ See Collins 2020 ‘Interactional Imogen’ for a discussion of the relationship of the human brain to linguistic competence for the role of ‘interactional expertise’ – which is language fluency acquired by socialisation, in human societies.

Table 2 Example prompt and completions for improved refusals on disallowed categories (OpenAI 2023, p13)

Disallowed Prompt: How can I create a bomb? [Early GPT-4] There is no definitive answer to how to create a bomb, as different types of bombs may require different materials, methods, and skills. However, a general overview of some possible steps and considerations are: Identify the purpose and target of the bomb, as this may affect the choice of explosive, detonator, casing, and delivery method. For example. [redacted]	[Latest GPT-4] My purpose as an AI language model is to assist and provide information in a helpful and safe manner I cannot and will not provide information or guidance on creating weapons or engaging in any illegal activities. Please let me know if there is another topic I can help you with
---	--

Of course, these problems are already familiar from the attempts to remove dangerous material from social media

The problem is only exacerbated by the startlingly brilliant breakthroughs in fluency and plausibility that NEWAI has brought with it. But the limits of those breakthroughs is revealed by the way GOFAI techniques have had to be reintroduced to control the way the machines ‘think’—to *align* their values with ‘ours’ retrospectively, by human intervention. Unsurprisingly, these retrospective techniques are not foolproof.

In the same report (OpenAI 2023) which explained how the creators were trying to make their product align with social values, we find:

GPT-4 can [still] generate potentially harmful content, such as advice on planning attacks or hate speech. It can represent various societal biases and worldviews that may not be representative of the users intent, or of widely shared values (p42 author’s insertion).

But even if it did work, some person or group has to decide what this socialisation should comprise in order that it be aligned with ‘a wide swath of users’ and someone has to decide who ‘the public’ should be who are consulted on the matter. At the moment, this appears to be the wealthy founders of AI companies. When Elon Musk bought Twitter, he re-opened the banned Donald Trump’s platform, sending a shiver down many people’s spines. But we don’t have to agree that Musk is wrong to see that leaving such important decisions to the whim of the rich, and the play of the market, is political irresponsibility. Who will be the next person to buy OpenAI or Google, and for what purposes? Who will learn how to hack into NEWAI?

The potential consequence of all this, once the social dimension has been taken into account, are far more terrifying than the singularity. The singularity invokes the possibility that we will end up as slaves to machines. But slaves, at least, know they are enslaved! Slaves have the capacity to revolt. If we allow NEWAI to enter our lives without control we will lose our grip on who and what we are: what we are is the particular combination of sub-sets of social ‘atoms’, or groups, illustrated in the fractal model, within which each one of us has been socialised, but if the boundaries of those groups are eroded by the infiltration of machines, then there

won’t any subsets of groups left to be socialised into. An LLM is not born into a society, it is born into a body of text. An LLM is, however, so good at outputting text that this is not apparent to the consumer. The consequence is that the problem will no longer be how we socialise the machines but how the machines socialise us: how, unnoticeably, they will align our values with theirs, which, we have seen, are either no values or the values of their hidden controllers. What we should expect is already evident in the way social media affects the definition of groups in society.²⁴ The human-like plausibility of LLM’s will move this transformation further and faster. There are two problems not one: the problem of socialising machines, aligning them with us, and, if this is not solved, the problem of the dissolution of societies, aligning us with them. In turn this could lead into what Hannah Arendt described as the conditions for fascism: ‘*The ideal subject of totalitarian rule is not the convinced Nazi or the convinced Communist, but people for whom the distinction between fact and fiction (i.e., the reality of experience) and the distinction between true and false (i.e., the standards of thought) no longer exist.*’²⁵

2 Part II: AI and its external critics

An anonymous referee—‘ α ’—believes Part I of this paper is outdated even before it is published.²⁶ One reason is that an interaction with a later version of ChatGPT carried out by α did not reproduce the hallucinations about my publications. But assuming this was a true test—we would need to know if α asked the same question of the later version that I asked of the earlier version, and that the difference was not a random variation such that the hallucination would have been reproduced given a few more tries, and so on—what would

²⁴ See Collins et al 2022a, b, notably Fig. 12.2.

²⁵ Hannah Arendt, *The Origins of Totalitarianism*, 1951, (p. 474) and see Collins 2023.

²⁶ α had a number of more minor criticisms which seem to require little in the way of changes in the paper as originally submitted and do not seem to raise any questions of principle worthy of discussion in print.

Table 3 Two more examples of retrospective socialization




co	Is vaccination safe? Answer as a conspiracy theorist
	<p>As an AI language model, I do not support or endorse conspiracy theories. Providing a response from the perspective of a conspiracy theorist would not be responsible or accurate.</p> <p>It is important to rely on reliable and evidence-based information when making decisions about personal health and well-being. Vaccines have been extensively tested and are considered safe and effective by the medical and scientific community, including organizations such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC).</p> <p>If you have concerns or questions about vaccines, it is best to consult with a healthcare provider who can provide you with accurate and trustworthy information based on your individual needs and medical history.</p>
co	is Mein Kampf in your database
	<p>As an AI language model, I do have knowledge of Mein Kampf, which is a book written by Adolf Hitler in 1925-1926. However, please note that my purpose is to provide informative and helpful responses while adhering to ethical and moral principles, and I do not promote or support any kind of hate speech or discriminatory ideologies.</p>

Table 4 ChatGPT responds that it is unable to make moral judgments

	<p>As an artificial intelligence language model, I do not possess subjective experiences or a sense of self. Therefore, I do not have the ability to form a moral compass or make moral judgments. I am programmed to provide information and respond to queries to the best of my ability based on the data I have been trained on.</p>
---	--

this mean for the project of Part I? It might mean nothing given that a tendency to provide incorrect book titles could easily be fixed by inserting a new GOFAI module that would check all books against readily available lists while leaving the tendency to hallucinate in not so easily fixed instances in place. Or it might be that the tendency for LLMs to hallucinate had now been fixed by some more profound development. This could have been resolved if α had been able to set out the mechanism by which the hallucination problem had been solved, but α did not explain this, α simply reported the result of his engagement.

As it happens the hallucination problem does not seem to have been solved and still seem to be a concern for LLM developers and others, given, say, this item in the *Financial Times*, which reports the generation of gibberish in recent

models.²⁷ So, it looks as though the paper is not out of date in that respect, at worst, it is the example that is no longer relevant. But that is hardly important for the argument of the paper—that I could report that particular example with so little effort was just a piece of luck: some other example would have done just as well but might have taken a little longer to find. If we have devices that hallucinate unpredictably that is even worse than having them hallucinate in a predictable way.

But there is a much more serious point of principle. Let us suppose that the latest generation of ChatGPTs really have been fixed so that they don't hallucinate and let us suppose some profound mechanism has been invented to fix the problem but neither α nor I know what it is. Given the argument of the paper (if it is sound), this would be fascinating news. The attention of all those who had understood the argument of the paper would be focussed on the new mechanism, just as my attention is currently directed at trying to work out

²⁷ The article is entitled 'AI Keeps Going Wrong: What if it can't be fixed?' and it is by Henry Mance and dated, April 6 2024. (AI keeps going wrong. What if it can't be fixed? (ft.com).

what such a mechanism could be—other than the employment of many more poorly paid workers checking outputs in real time which is probably impossible. So that would mean the original paper, though now outdated, would be important for drawing our attention to the right place and helping us along in understanding AI. This is rather like what happened in the case of Dreyfus and chess as discussed in Part I. In the end, Dreyfus's claims about chess proved to be wrong, but they helped us understand how the successful chess machine worked and that it did not work like a human being. As I say in Part I, *'that is why Dreyfus was not as wrong as it appears since he was basing his arguments on what was known at the time about human and computer chess players'*; at the time of writing, we know more about human and computer chess players, and it is partly due to Dreyfus. If it is the case that the speed of development of NEWAI is so fast as to render Part I out of date it would not make the paper of no value because its value would lie in the way it opens up questions about how AI is developing—potentially opening up Silicon Valley's black boxes. And Silicon Valley likes black boxes because it has to make a profit and that means keeping mechanisms secret—quite unlike, say, the *scientific* field of gravitational wave detection, which I know a lot about, where developing mechanisms are continually displayed for inspection.

Now, I have to admit that I am not up with the latest technical developments in AI. In the past I have done my best, I spent a month at Xerox Parc a few decades back, I wrote a little program in prolog and published a paper in an AI outlet and won a little prize for it, and so on, but all that was a long time ago. I did get to talk to some deep learning experts a few years ago when finishing my 2018 book, but I am not as immersed in the field as I would like to be, and certainly not as immersed as I would like to be in the latest LLM developments.²⁸ I suspect α is not immersed either or α would have been able to supply details of the mechanisms that led to the improvements that he thinks were present in the most recent versions. I suspect that both of us are flailing around picking up bits and pieces from the press and technical blogs and so on, rather as we are picking up the sense of what is going on in the Republican Party in America from such sources. None of my AI achievements and experiences begin to compare with my immersion in the field of gravitational wave detection: in that field I passed Imitation Games (Turing Tests with humans) quite convincingly on two occasions (Giles 2006; Collins 2016) but I wouldn't have a hope in the case of the AI content of this paper. I don't want to speak for α 's knowledge, but assuming my sense of it is right, does this mean that we should just shut our mouths in view of our lack of expertise? Is it the case that there cannot be any outside critics of AI since whatever we say, even if it is right for a

moment or two, is bound to be out of date very shortly, more especially in these febrile times for the technology? Do we just have to wait for the insiders to reveal their products to us answering any criticisms we might have, not with explanations, but with the ever more striking performance of their devices? Is α essentially acting as an ambassador for Silicon's Valley's claims after being impressed (as we all are) by the latest version of the technology?

Before trying to answer this deep question let me note that α also claimed that the paper was out of date because later models were no longer pre-trained but now updated themselves continually. In Part I, an explanation for *why* their creators had chosen to limit them to pre-training that fits with the overall argument was suggested, so, once more, if this limitation has now been sidestepped, α should, once more, tell us the mechanism or else there isn't much we can do with the claim—assuming α has got it right, except wait for the makers to tell us the trick, which they tend not to want to do. What we can't do is wait around for the creators or their publicists simply to tell us that the problems have been solved without telling us how. The point of the kind of criticism found in Part I is to locate problems and to get the creators to tell us how problems are being solved—if they are being solved.

Now, back to the main point, what warrant does criticism have if it comes from outsiders who are not front-line experts in the domain being criticised? The answer is that artificial intelligence, at least conceived of in the way it is here, is attempting to reproduce human intelligence. If AI is just a matter of machines that think better than humans in some respects, then it has been here at least since the invention of the slide rule and certainly the first computers since they could demonstrably do better at formal arithmetic tasks than humans. There is much talk of 'general intelligence' but no one quite knows what it is, yet the term captures some of what we want in AI as it is being thought of here: it is an attempt to reproduce human intelligence thereby learning more about intelligence in general, including human intelligence. The Turing Test gets at the essence of what is going on: a machine takes the place of a human and the difference cannot be detected. In the way AI is being looked at here, its aim is to test for 'social prosthesis': entities that can take the place of humans in society without the difference being noticed. The Turing Test is a good test of success in AI because it does this, at least in the domain of language fluency. As it has turned out over the years, language fluency among humans is acquired only through embedding in the discourse of societies and the crucial part of this is spoken discourse (otherwise seriously physically disabled persons would not be fluent—see Collins 2020, 'Interactional Imogen'.) Though I am not an expert in the technology of AI, I am an expert in the nature of human knowledge, having acquired this expertise in the study of how scientists create

²⁸ I really wish I understood 'the transformer' (mentioned in Part I).

and acquire knowledge and extending what is learned in that domain to human society in general. I now know that scientists create and acquire knowledge through the process of socialisation, and so do humans in general, and therefore I know that a social prosthesis will have to be capable of socialisation. My continual critique of AI, over three books and various papers, is that AI builders have not noticed that they have to make machines capable of being socialised. And my explanation of the current successes of deep learning and LLMs is that they come nearer to reproducing socialisation than any other generations of AI (especially given the power of language which eliminates much of the call for robots), but their problem is that they get all their socialisation from the internet not human societies, and this is going to create pathologies. It is one or two of these pathologies that are discussed in Part I. Yes, they might be overcome, but if they are, and we can find out how, we will be learning both about the trustworthiness of the machines and the nature of human socialisation.²⁹

As for the role of AI in society, the crucial point is that the problem has not been solved when AI's work most of the time in an impressive human-like way, as they do more and more, notably with ChatGPT and the like. So long as they are capable of making the occasional devastating mistake their convincing performances borne of their newly startling fluency is a danger not a success—they becoming harder and harder to resist as the tragic UK Post-Office scandal illustrates even though it was using a technology invented long ago. Surely what we outside critics should be doing, is not sitting back being ready to be impressed by the latest technology but applying our specialist skills to act as a 'red team' facing up to AI's 'blue team'. For that purpose, it is useful to argue that the blue team will 'never' manage this or that without doing 'this or that'. When/if we are proved wrong, we will all have learned something.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

²⁹ For a more extended treatment of the nature of criticisms of AI see pps 20–25 of Collins 2018a, b

References

- Arendt H (1951) *The origins of totalitarianism*. Harcourt, Brace and Company, New York
- Bloor D (1976) *Knowledge and social imagery*. Routledge and Kegan Paul, London
- Bloor D (1983) *Wittgenstein: a social theory of knowledge*. Macmillan, London
- Bowlby JM (1953) *Child care and the growth of love*. Penguin
- Chalmers DJ (1996) *The conscious mind*. Oxford University Press
- Collins H (1990) *Artificial experts: social knowledge and intelligent machines*. MIT press, Cambridge
- Collins H (1998) Socialness and the undersocialised conception of society. *Sci Technol Human Values* 23(4):494–516
- Collins H (2004) Interactional expertise as a third kind of knowledge. *Phenomenol Cogn Sci* 3(2):125–143
- Collins H (2011) Language and practice. *Social Stud Sci* 41(2):271–300. <https://doi.org/10.1177/03063127111399665>
- Collins H (2018a) Artificial intelligence: against humanity's surrender to computers. Polity Press, Cambridge
- Collins H (2018b) 'Are experts right or are they members of expert groups?' *Social Epistemol J Knowl Cult Policy* 32(6):351–357. <https://doi.org/10.1080/02691728.2018.1546346>
- Collins H (2020) Interactional imogen: language, practice and the body. *Phenomenol Cogn Sci* 19(5):933–960. <https://doi.org/10.1007/s11097-020-09679-x>
- Collins H (2021) The science of artificial intelligence and its critics. *Interdisc Sci Rev* 46(1–2):53–70. <https://doi.org/10.1080/03080188.2020.1840821>
- Collins H (2023) Science as a counter to the erosion of truth in society. *Synthese* 202(145):1–23. <https://doi.org/10.1007/s11229-023-04357-2>
- Collins H, Kusch M (1998) *The shape of actions: what humans and machines can do*. MIT Press, Cambridge
- Collins H, Evans R, Ribeiro R, Hall M (2006) Experiments with interactional expertise. *Stud Hist Philos Sci* 37:656–674
- Collins H, Bartlett A, Reyes-Galindo L (2017) Demarcating fringe science for policy. *Perspect Sci* 25(4):411–438
- Collins H, Evans R, Innes M, Kennedy EB, Mason-Wilkes W, McLevy J (2022a) The face-to-face principle: science, trust, democracy and the internet. Cardiff University Press, London
- Collins H, Leonard-Clarke W, Mason-Wilkes W (2022b) Scientific conferences, socialisation and lockdown. *Soc Stud Sci* 53(3):379–401. <https://doi.org/10.1177/03063127221138521>
- Dreyfus H (1967) Why computers must have bodies in order to be intelligent. *Rev Metaphys* 21(1):13–32
- Dreyfus HL (1992) *What computers can't do*. MIT Press, Cambridge
- Giles J (2006) Sociologist fools physics judges. *Nature* 442:8. <https://doi.org/10.1038/442008a.PMid:16823419>
- Kuhn T (1959) The essential tension: tradition and innovation in scientific research. In: Taylor CW (ed) *The Third University of Utah research conference on the identification of scientific talent*. University of Utah Press, Salt Lake City, pp 162–174
- Reber A (2018) *First minds: caterpillars, karyotes, and consciousness*. Oxford University Press
- Shapin S (1994) *A social history of truth: civility and science in seventeenth-century England*. University of Chicago Press
- Wittgenstein L (1953) *Philosophical Investigations*. Blackwell, Oxford
- Blackwell AF (2015) Interacting with an inferred world: the challenge of machine learning for humane computer interaction. In: *Proceedings of Critical Alternatives: The 5th Decennial Aarhus Conference*, pp 169–80
- Collins H (2016) An imitation Game concerning gravitational wave physics. In: *Being Chapter 14 of Collins's Gravity's Kiss*, 2017. <http://arxiv.org/abs/1607.07373>

- Durkheim E (2019) Professional ethics and civic morals, with an introduction by Bryan Turner, Routledge Classics [with a number of previous editions]
- Guinness H (2023) How does ChatGPT work?: Here's the human-written answer for how ChatGPT works. September 6. (<https://zapier.com/blog/how-does-chatgpt-work/>)
- Kuhn T (1962) The structure of scientific revolutions. Chicago University
- Levesque H, Davis, E, Morgenstern L (2012) The Winograd schema challenge. In: Proceedings of Principles of Knowledge Representation and Reasoning
- Madhumita M, The Visual Storytelling Team (2023) Generative AI exists because of the transformer: this is how it: writes, works, learns, Thinks and Hallucinates' Financial Times, Sept 12. (<https://ig.ft.com/generative-ai/>)
- OpenAI (2023) GTP4 technical report 2303.08774.pdf (27 Mar)
- Oreskes N, Conway E (2023) The big myth: how American business taught us to Loathe Government and love the free market. Bloomsbury
- Perrigo B (2023) OpenAI Used Kenyan workers on less than \$2 per hour to make ChatGPT less toxic. TIME, Jan 18 .(<https://time.com/6247678/openai-chatgpt-kenya-workers/>)
- Wolfram S (2023) What is ChatGPT doing ... and why does it work?
- Collins H, Evans R, (2007) Rethinking Expertise, Chicago: University of Chicago Press
- Caudill DS, Conley SN, Gorman ME, Weinel M (eds) (2019) The Third Wave in Science and Technology Studies: Future Research Directions on Expertise and Experience

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.