19

Artificial General Intelligence

Shocks, Sentience, and Moral Status

Peter Millican

The ambition to create an *Artificial General Intelligence* or *AGI* has been around at least since Alan Turing introduced his famous test for machine thinking in his 1950 paper 'Computing Machinery and Intelligence'.[1] This test was based on the ability to simulate human-style textual conversation, potentially over a wide range of topics, so as to be indistinguishable from a real person. But subsequent attempts to pass that test were, until very recently, confined to relatively crude chatbots following the style of Joseph Weizenbaum's *ELIZA* of 1966, which emitted ready-prepared outputs in response to specific prompts (so, for example, mention of one's mother might meet with the response, 'Tell me more about your family'; and 'I need X' with 'What would it mean to you if you got X?'). Such chatbots were interesting toys, but could not plausibly be described as genuinely *intelligent*. That situation has changed radically with the arrival of *ChatGPT*, which many people see as potentially leading to a genuine AGI. And this is not only of theoretical interest, because it raises a host of moral questions, including whether such an AGI would itself be worthy of moral consideration.

The Deep Learning Revolution

Although serious research in AI had previously been fruitful in many different directions, the actual achievement of 'general' AI seemed very distant until around 2015. Then a new program, *AlphaGo*—developed by the company *DeepMind* using 'deep learning'

techniques—burst onto the scene, defeating European champion Fan Hui 5-0 in a match of Go, a game which until then had been considered too subtle and complicated for computer algorithms to master in the foreseeable future.[2] The following year, *AlphaGo* sensationally went on to defeat World Go Champion Lee Sedol 4–1. And in 2017, a more generalised program, *AlphaZero*, which was capable of teaching itself new games—without any input from human experts or game databases—was able to defeat the world champion chess program *Stockfish* after only a few hours of self-training (i.e. playing games against itself). This was 20 years after *Deep Blue* had defeated Garry Kasparov, then human world champion and arguably the strongest chess player of all time. But programs such as *Deep Blue* and *Stockfish* had been specifically programmed to play chess, with input from expert players and using algorithms that were finely tuned to reflect established theory. *AlphaZero* was an altogether more impressive system, apparently able to learn new skills 'from scratch', and thus representing massive and potentially frightening progress towards Artificial General Intelligence.

Part of the promise and threat of deep learning lies precisely in its ability to represent all kinds of information in ways that it works out for itself in response to training data and feedback. This information is stored implicitly within the links of a neural network that consists of layers of artificial neurons. The input layer of the network is set up to reflect the specific problem case, so for example the activation level of the input neurons might represent the state of the squares in a chessboard position (which is to be assessed), or the colour of the pixels in a digital image (which is to be classified). The output layer is set up to signal the corresponding solution, respectively the chess assessment (e.g. how good the current position is for White) or the image label (e.g. 'cat' or 'dog' if the task is to classify images of pets). But in a deep network there may be many intermediate layers, with patterns of activation passing down through the layers from the input layer to the output layer. Beyond the input layer, the activation of each neuron depends on the inputs that it receives from the neurons to which it is connected in the previous layer. These inputs depend both on the level of activation of those previous neurons, but also—crucially—on the weight given

to the relevant connection. It is these weights that are adjusted during the learning process, which typically involves going iteratively through the training data, assessing the resulting outputs and gradually refining the relevant weights until a sufficient match between inputs and outputs has been achieved. Neither the weights, nor the roles of the neurons in the intermediate layers, are predetermined when the learning process starts. By the end, the immensely complex pattern of weights implicitly represents what the network has learned, but in a way that unaided humans will find impossible to interpret, and whose behaviour they will be able to predict only inductively (i.e. by experience).

All this can look almost magical, enabling a deep network to solve problems that we have no idea how to address, and using calculations that are beyond our comprehension even when they have been discovered. But we should not be misled to suppose that the machines themselves 'understand' what they are doing in any reflective way—and not only because they are completely non-sentient (and hence have no awareness or conscious understanding of anything). For their way of working is far more closely analogous to our own unconscious pattern-recognition than it is to how we think when explicitly calculating or reasoning about something. And we are unlikely to consider that we really understand *how* we ourselves operate when recognising an animal as a dog, or a symbol as the letter 'f'—we too learn these things by examples, building up neural structures of which we are completely unaware, and whose results we find hard to articulate. Likewise in expert chess-playing—often considered a paradigm of explicit calculation—many moves are decided by a sort of trained instinct, involving implicit recognition of a position's characteristics and appropriate strategies, to the extent that a grandmaster will often choose a move with negligible calculation, and if asked, may initially have little more to say than 'in this sort of position, that is the right move to play'. (He could probably expand on the strategic significance of the move given longer to think about it, but that deeper reflection would not have featured in his move choice within the game.) There is an obvious analogy here to a neural network's method of position evaluation by pattern recognition. Of course, the grandmaster also needs to be

able to tell when explicit calculation is needed, for example, when tactical combinations are imminent, but *AlphaZero* too can perform explicit calculation, exploring options down the 'game tree' of branching possibilities. This tree-searching ability was built into it, along with the rules of the game—the initial position, how pieces move, what counts as a win, draw, or loss etc. So in this sense it did not really teach itself chess *entirely* 'from scratch', despite its lack of access to human chess experts. Within that closed, rule-governed context, it was able to generate its own training data and feedback by playing lots of different games against itself, learning *by experience* which patterns were most conducive to success.

The Practical Shock of *ChatGPT*

Since 2022, we have witnessed a new 'AI shock' which seems even more significant and widespread than any that has come before, with the release of *OpenAI*'s chatbot system *ChatGPT*. This again has been developed using techniques of deep learning, but this time with colossal human textual input in the form of around 300 billion words taken from various sources on the internet. And now we suddenly—and unexpectedly—are faced with a technology that seemingly has the potential to pass the 'Turing Test' in its full generality: to converse plausibly, flexibly, coherently, and informatively about a vast range of topics, and without relying on pre-prepared outputs.

The arrival of *ChatGPT* raises major foundational questions about AGI, to be discussed below. But first, we should note some serious practical worries, with significant moral implications that would deserve extended discussion in themselves. The answers that *ChatGPT* generates look very plausible as humanly composed text, but they are unreliable, not fact-checked, and frequently involve 'hallucination' (i.e. invented claims). They are also likely to reflect biases in the source texts, and can be deliberately manipulated by asking *ChatGPT* to produce arguments for falsehoods that serve one's purposes. The system's ability to turn out believable and varied disinformation and propaganda, almost instantly and at minimal

cost, provides obvious opportunities for those wishing to manipulate the online environment, for example by flooding discussion or review sites with submissions that all support a particular point of view, and yet give the appearance of having been written independently.

Related concerns arise from *ChatGPT*'s ability to generate text in specific styles as requested by the user, and thus to impersonate prominent individuals or organisations, bringing risks of fraud and deceit. That ability can be exploited more positively in the drafting of standardised forms of text, for example legal documents, though this raises serious concern about potential job losses, as humans are replaced in various routine capacities. Such replacement could in turn lead to an over-dependence on automated systems and a lack of critical perusal as the relevant processes—optimised for cost rather than quality—cease to involve reflective and suitably trained humans. This could be economically devastating for entire categories of human workers, the same workers whose online contributions over the years have been plundered by *ChatGPT* to produce its indirectly plagiarised outputs. Such developments also risk undermining accountability, as important decisions are effectively outsourced to an automated system whose workings are inscrutable and subject to numerous other concerns (as just explained). Modern companies, often intensely competing with their rivals, are systematically incentivised to reduce costs and increase revenue, and they are typically partitioned into units that also focus individually on their own budget and bottom line, with little role within these processes for broader societal concerns. So there is a significant danger that even companies that ostensibly have the best of intentions, and with employees who share those ideals, will end up following paths that optimise their short-term competitive and financial benefits, while sleepwalking progressively into a situation which nobody would have wished for.

Is *ChatGPT* an AGI?

Putting these practical concerns to one side, how should we assess *ChatGPT* and its (fast multiplying) cousins from a theoretical point

of view? The system was developed by applying statistical analysis on those 300 billion words of textual data to create a *Large Language Model* (LLM) which—to simplify somewhat—records the probability that any given sequence of words (within its broader context) will be continued in different ways. And accordingly, *ChatGPT*'s primary method of working is to predict which individual words are most likely to follow in any particular textual context, and then to choose one of these words (but not always the most likely word, since an element of randomness enables it to respond differently if the same context is repeated). That chosen word is then added to the existing text, and the next word chosen in the same way, and so on. Alongside the purely automated learning which generated the (reportedly) trillion or so parameters—that is, stored probabilities—in the LLM (reflected in the deep network weights, as explained earlier), human users improved the system's responses through supervised fine tuning, in which it was given a wide range of typical prompts (i.e. potential user inputs), together with suitable human-crafted responses. Then a stage of reinforcement learning from human feedback was applied, whereby the system generated a range of responses for each given prompt, and humans assessed the relative suitability of those responses. This feedback was then statistically analysed to generate a *reward model* (i.e. a rough method of calculating the quality of response outputs), which could in turn be used to give positive or negative feedback to the network, thus refining its selection of appropriate responses.

This is, of course, a hugely simplified description of how *ChatGPT* was created and how it works. But two points here are particularly worthy of emphasis. First, its method of working is based entirely on imitating—with variation—the sorts of responses that occurred in the massive textual resources that were used to train it. Secondly, the information which it stores implicitly within its trillion or so network weights has been tuned to reflect *the characteristics of the textual data*, rather than *the characteristics of whatever domain the text might concern*. Hence, for example, if you play chess against *ChatGPT* using a common opening (e.g. a main line Sicilian Defence), it will initially respond with some of the same sensible moves that dominate its training data, and thus give the impression

that it understands what is happening on the chessboard. But it stores no internal model of the board position, has no mechanisms of analysis or 'lookahead' (unlike *AlphaZero*), and is therefore subject to absurd errors, such as overlooking obvious captures, or allowing your bishop to capture one of its own pieces even when the diagonal between them is blocked. Play a less common opening, and the problems will appear much more quickly (e.g. losing track of a pawn on the fourth move of a King's Gambit). Ask it to solve a simple chess puzzle—for example, to devise a position in which White's king is on h1, Black's king on h8, and White, with only one further piece on the board, is able to checkmate Black by playing that bishop to b2—and it will have no clue. Perhaps before long these particular foibles will be dealt with, by linking up *ChatGPT* to a dedicated chess engine that really does have a model of what is happening on the board together with relevant analytical algorithms. My key point here, however, is that its apparent *general intelligence*—its ability to converse 'intelligently' about a vast range of topics where it has not been augmented with specific assistance—is based on the illusion that it has some internal understanding of what is being discussed. But unless the language model in some domain is able to generate indirectly a reliable model of the reality (which looks most plausible in domains that are primarily conceptual or expressive), it will have nothing like such internal understanding, and can quite appropriately be described as a 'stochastic parrot'.[3]

The hype surrounding *ChatGPT*, and failure to appreciate the points just made, could lead to real dangers if humans come to depend on its outputs by assuming that they have been generated through some kind of genuine understanding. One obvious area of concern is software development, where the apparent ability of *ChatGPT* to perform impressively in programming exercises might tempt companies to rely on it for code production, and thus hugely economise on their software engineering teams. But as in routine chess openings, the reason why *ChatGPT* does so well in those exercises is not that it rigorously analyses some internal model of what is going on in the code that it generates, but rather, that its own implicit textual model is based on analysis of the many

millions of programming examples that it has been given, and it accordingly generates the code by mimicking, with what appear to be appropriate textual variations, the dominant patterns. So if you have a programming task that is very standard, or which deviates from a standard task in relatively common ways (e.g. where teachers have devised non-standard programming tasks for their students, such as sorting an array of numbers but with odds preceding evens), then *ChatGPT* might well come out with a correct solution. But if the task involves any nuances (or combinations of factors) that are less common or entirely novel, then there is a serious risk that you will end up with code which has the double disadvantage of looking very plausible—and quite possibly works well in most cases—whilst actually being incorrect, thus making the errors especially hard to identify.

None of this is to deny that in expert hands, *ChatGPT* can be a useful programming assistant, providing quick and targeted search through the mountains of online programming resources to generate 'first draft' answers that are often on target. But the use of *ChatGPT* in programming by those who are unaware of its foibles, or insufficiently expert to check carefully what it produces, carries the risk of generating false confidence, both in the code that it generates and in the programmers who rely on it (whose lack of understanding could be masked by their success in turning out plausible code). Empirical studies consistently show that far more programmer time is taken fixing 'bugs' than writing original code, so focusing on saving time on the latter, at the cost of code quality, is a dangerous false economy. Companies whose managers are taken in by the hype of supposed AGI could easily find themselves relying on large amounts of parroted code that reflects limited understanding of the problem at hand, and which, when it goes wrong, is either a nightmare to debug or needs to be replaced entirely. This is a serious potential danger, because the combination of poor software quality and excessive confidence in its reliability has notoriously led to many public scandals, involving huge personal and commercial loss (for example, the Horizon IT scandal which led to numerous unjustified prosecutions of British sub-postmasters, resulting in wrongful imprisonment, bankruptcy, and even suicide).

Could an AI System be Sentient or Worthy of Moral Consideration?

Despite all these reservations and limitations, it is understandable that the development of Large Language Models has triggered huge interest and concern around the possibility of AGI. *ChatGPT*'s ability to converse so plausibly over such a wide range has come as a tremendous surprise to most people, including AI researchers, and the long-standing influence of the Turing Test has led many to view such conversational competence and versatility as the hallmark of AGI.

Let us, then, imagine that in the not-too-distant future, some comprehensive AI program is designed using broadly similar technologies, enhanced where necessary (as has already started happening) with a wide range of specialist modules that enable it to handle tasks requiring particular types of detailed calculation that go beyond the powers that it has managed to learn for itself. In these cases, the conversational interface of this hypothetical program translates natural language queries into some internal form which explicitly identifies the required task. Then if the task involves chess, a dedicated chess engine will be used to find the answer, which is then translated back into natural language. Likewise, if the task involves mathematics, or programming, or whatever, the necessary processing will be done either by a specialist module, or perhaps by some more versatile system which is able to construct and manipulate representations of non-standard scenarios (e.g. to deal with arbitrary logic puzzles). In 2024, this possibility does not seem nearly as remote as it did a decade ago.

Suppose, then, that we are presented with this impressively versatile system, which seems to be capable of plausibly intelligent responses over a vast range of topics, and which—unlike *ChatGPT* and its ilk—also appears to be highly reliable over the entire range, generating answers that are overwhelmingly correct.[4] We would thus have a system which plausibly passes the Turing Test and, in at least one reasonable sense of the phrase, is an *Artificial General Intelligence*. Does this mean, however, that we would

also have an AGI in another common sense of that phrase, namely, a system capable of human-level cognition in general, including that mysterious, elusive, and morally fundamental quality of sentience or consciousness? The two senses are very often confused or merged, but I shall end by arguing that they should be sharply distinguished.

Mostly, when we encounter two systems whose external behaviour is similar, we naturally and reasonably suppose that they are likely to be similar internally also. So when we see animals acting in ways that are analogous to our own primitive behaviour—for example seeking food or shelter, defending themselves, nurturing infants, or interacting with a potential mate—we suppose that they are intentionally motivated in much the same way as ourselves, either by instinctive desires or by the prospect of pleasures and pains. And if their behaviour is highly sophisticated and flexible in response to variable circumstances and interactions with other animals, we would also suppose that this motivation is to some extent controlled by conscious reasoning. We would presumably make similar judgements if we were to encounter some apparently intelligent alien life form, and it might then seem inevitable that, as Turing advocates, 'fair play for the machines' mandates our applying the same criteria to artificial intelligence systems.[5]

But to illustrate where this seductive line of thinking goes wrong, suppose that we have standard criteria for identifying a wood fire in terms of its external signs: patterns of heat and light, crackling sounds, and smoke. And now suppose that we are confronted with an electronic device which is intricately designed to mimic all of these features, using heating elements, lights, speakers, and artificial smoke. Knowing how it works, do we have good reason to apply our standard criteria and conclude that wood burning is taking place? Of course we don't, because we know that those standard criteria are being fulfilled in a *different* way using clever electronics rather than real combustion, and we are certain that no wood at all is involved in the process.

A parallel point can be made about our supposed AGI, for we know at least in outline how it is generating its 'intelligent'

responses, using sophisticated electronics implementing deep neural network structures, fed with a massive textual corpus and tuned by clever machine-learning algorithms that have been designed precisely for the purpose. We might well be somewhat mystified as to how exactly the finished system works in terms of all those inscrutable representations stored implicitly in the network weights, but we know in general terms how the network is operating—including a *precise* understanding of the functions of its individual digital components—and we know that all this involves no element whatever of consciousness or sentience on the part of the machine. So to suppose that somehow the system magically becomes conscious once some threshold of complexity has been crossed (in the way beloved of sci-fi films such as *Terminator*) is complete fantasy.

But if this argument works, a critic might ask, why would it not apply equally to other humans, albeit they are made of *biological* (rather than *electronic*) hardware? And would that not lead to solipsism (as Turing alleged in his 1950 paper, §6.4)? One obvious reply is that we emphatically *do not* have a complete functional understanding of our own biological hardware: brain science is at a very early stage, and its biochemistry is extraordinarily complex (not to mention that our understanding of basic physics remains fundamentally incomplete). So we have no good reason for assuming that the behaviour of our neurons can be abstractly modelled in a similar way to digital components, let alone that they could adequately be replaced with electronic substitutes (as is commonly assumed in fanciful thought-experiments). Another obvious reply is that we are immediately aware of our own consciousness as something which is more than merely an abstraction, and is causally efficacious (e.g. we seek things that give us conscious pleasure, and shun things that give us pain). This causal efficacy, indeed, is key to its evident evolutionary tuning, whereby conscious pleasures and pains are strongly associated, respectively, with things that are evolutionarily beneficial and harmful (a coincidence which would otherwise be wildly improbable). This gives an excellent basis for both inferring consciousness in other beings that have evolved in a similar manner,

and for asserting that our own mental processes cannot be reducible without remainder to the digital functions of logic gates etc., which may combine to generate sophisticated *abstract* processing (such as calculation of chess moves), but cannot generate a novel *causal* mechanism that goes beyond those functions.[6] We do not know how our consciousness is caused by our biological hardware (or, some would say, by an immaterial soul), but however it may be caused, we can be confident that it is not purely abstract.

This case against the possibility of conscious digital machines is not particularly sophisticated, and may appear to many as plain common sense. But I suspect that it is often overlooked precisely because of the unfortunate conceptual blurring implicit in our notion of 'intelligence'. On the one hand, we tend to judge intelligence in terms of efficient and effective information processing in the solution of relatively complex problems. On the other hand, we tend to presume that intelligent entities will be purposively motivated by conscious goals, devoting their energies to solving the sorts of problems that stand in the way of their achieving those goals. Since Alan Turing's momentous invention of his 'computing machines' in 1936, however, we know that these two aspects can come apart: it is entirely possible to have efficient and effective information processing without any trace of sentience, and it is likewise possible to have systems—for example chess computers—that behave purposively to achieve goals without those goals being the least bit conscious. Thus it is perfectly reasonable to allow that such systems are exhibiting genuine intelligence, in the sense of *sophisticated information processing to good effect*, without in any way having to accept that this implies genuine *sentience*. Indeed, we have excellent reason for denying that the latter is even possible for our artificial electronic machines, because we know (in general terms) how they operate, and this removes any justification for such fanciful speculation. So even while our own conscious awareness and its biological basis continue to be fundamentally mysterious to us, this gives no reason for extending the scope of that mystery so that we come to wonder about the potential sentience of our electronic computer systems. Their non-sentience is something of

which we can be certain, precisely because we know how they operate right down to the digital functions of their most basic computational components, and we also know how those functions give rise to the intelligent behaviour. This is an argument based on positive knowledge, not ignorance.

As for whether such artificial systems are worthy of moral consideration (i.e. as moral *patients* rather than *agents*), my answer would be an emphatic 'No!' Philosophers such as Immanuel Kant may have argued that moral value is a matter of *rationality*, but that always seemed implausible to those of us who took animals (and Darwin) seriously—it is far more plausible to see moral consideration as applying to creatures that are capable of *feeling*, and here David Hume rather than Kant points the way. Indeed, this position is significantly strengthened by the creation of artificial intelligences which can perform rational information processing but lack any inner life. Utterly incapable of real desires, pleasures or pains, fulfilment or suffering, it follows that they also lack any intrinsic interests. Hence such artificial intelligences cannot possibly be worthy of moral consideration.

Acknowledgement

I am very grateful to Vince Conitzer, David Edmonds, Mark van der Wilk, and Michael Wooldridge for helpful comments on drafts of this paper.

Notes

1. In the journal *Mind* volume 59, pp. 433–60. Turing himself did not use the term *Artificial General Intelligence*, which started to become commonly used around the turn of the century.
2. For a vivid and accessible account of the history of machine learning and the beginning of the deep learning revolution, see pp. 167–200 of Michael Wooldridge, *The Road to Conscious Machines: The Story of AI* (Penguin, 2020).
3. The term 'stochastic parrot' was coined by Emily M. Bender in 2021. Recent work by my Oxford colleagues Philipp Koralus and Vincent Wang indicates that the most recent generations of *ChatGPT* have become better at mimicking

good human thinking, but also parroting *fallacious* human thinking (arXiv:2303.17276v1). Prompt engineering and generation of synthetic datasets might mitigate this weakness, but again, such methods focus on enabling the system to find answers through modelling of *human language*, rather than of *the relevant domain*.

4. Note that this might be very hard to establish. Even if the system responds brilliantly well in the large range of situations that have been anticipated (and perhaps "patched" with specialist modules), it could still remain the case that when something is asked which "falls between the cracks", the system will be utterly—and perhaps dangerously—useless. This contrasts with typical human performance, where we are able to infer with confidence that someone who is excellent in general will not be utterly stupid beyond their expertise. So we need to be very cautious about extrapolating such expectations to AI systems.
5. For appreciation of how far the Turing Test setup was motivated by a desire for such fairness—rather than primarily by seeing human indistinguishability as a *criterion*—see the article cited in the Further Reading, pp. 37–8, 45–6. See also pp. 47–9 for further discussion of the issues raised here.
6. This point applies even if we have an AI system which is itself 'evolving' algorithmically in response to feedback. All such 'evolution' involves specific changes in the operation of a system whose basic processes we understand well, and which we know to involve no element of consciousness. By contrast, it seems that biological evolution was able somehow to exploit properties of matter which we do not currently understand, and which generate distinct causal mechanisms associated with consciousness.

Further Reading

Peter Millican, 'Alan Turing and Human-Like Intelligence', in Stephen Muggleton and Nicholas Chater (eds), *Human-Like Machine Intelligence* (Oxford University Press, 2021), 24–51, especially pp. 36–49.