To be published in *Artificial Intelligence and its Discontents*, ed Colin Garvey, in series *Interdisciplinary Science Reviews*.

The *science* of artificial intelligence and its critics

Harry Collins

Introduction

The title of this volume, 'AI and its Discontents', says something about the AI domain which is reflected in the contributions. They are very different in method, resources, and motivation and this, I suggest, reflects AI's widely varying presentation of itself as a science. Therefore, let me explain that here I am trying to use my specialist expertise to contribute to AI *as a science* (even though I am not an AI specialist, narrowly conceived). I am addressing an AI which sees itself as an attempt to *explicate* the workings of human intelligence by creating 'mechanical causes and effects that mimic human actions'.[1] The goal hasn't been reached and may never be reached but we have already learned a lot about human intelligence in the attempt. Thus conceived, the science of AI contributes to my specialist subject, which is the nature of knowledge, and my specialist subject can contribute to the science of AI since it can explain what AI should be trying to do if it is to emulate human knowledge. So, in so far as what I say about AI is critical of some of its the claims, it is meant to be productively critical.

---

[1] According to me, this is the third of four meanings of 'to explicate' Collins, 2010, *Tacit and Explicit Knowledge,* Table 4, p 81

Productive and audience-directed criticism.

Criticism is a central feature of science as it generally thought about. Robert Merton argued that 'organised scepticism' was essential for science, and the same follows from Popper's claim that science was more about falsification rather than corroboration. What they were talking about was productive technical criticism. The pre-requisite for productive criticism is that there must be at least *a chance* that the criticised will learn from it and act on it. They probably won't because, as Max Planck is said to have said, 'science advances funeral by funeral' – that is, scientists don't change their minds much – but the productive critic has to *aspire* to change the insiders' minds. This means the critic must start from inside the position of the criticised and try to reveal its flaws in terms recognisable and valued by the criticised. The other kind of criticism is 'audience directed criticism' which aims not to convert the practitioners of the science but to persuade outsiders. Audience directed criticism is typically found in religion, or politics, or commerce, or the courtroom, where the aim is not to convince the technically savvy but to convince the ordinary person who is looking on from the outside: in religion you are not trying to convert another religion's priests but trying to sway a congregation; in politics you are not trying to convince the opposition party but the electorate; in commerce you are not trying to convince a competitor that your product is best but to convince consumers; in the courtroom the prosecution is not trying to convince the defence but the jury. In all such cases the criticism does not have to bear too strongly on what you are against; indeed, an effective way to win over outsiders is to misrepresent the opponent so as to make their position seem less credible than it would be if accurately portrayed. This is hopeless if you are aiming for productive criticism because no-one is

going to change their minds in response to criticism that starts by misrepresenting the target.

Productive criticism is much harder than audience-directed criticism because it must start with a journey into the heartlands of the opponent's world. In big sciences, such as high-energy physics, where all the experts tend to have been gathered into the research team, it is recognised that the sharpest and most useful criticism is likely to come from inside the team not outside. In high energy physics they solve the problem by setting up opposing groups within the overall organisation. The way this attitude showed itself in the case of the first detection of gravitational waves, which also recruited pretty-well all the experts there were, was that it took the insiders five months of feverish work from the appearance of a promising signal on September 14th 2015, to be satisfied that they had themselves critically examined every aspect of what they had done, and considered every reasonably imaginable flaw, before they announced the discovery; the press conferences were not held until February 11th 2016.[2] This tradition of productive of criticism is also found in AI but seems to be less dominant than it once was. It is represented by such as, Joseph Weizenbaum, Terry Winograd, Bert Dreyfus, and Lucy Suchman. This list is not meant to exclude anyone else who aspires to be part of it, including myself, of course.[3]

---

[2] The author watched the entire process from the inside: Collins, 2017.

[3] In a 'Postscript' to this piece I will discuss the very recent productive criticisms of Marcus and Davis. As Colin Garvey reminded me, the reception by the technical community of the critical remarks of some of those in this list illustrate Planck's point that the attempt to engage in constructive, technically informed, criticism is no guarantee of being listened to or even tolerated.

I am not arguing that the only worthwhile criticism of science has to be technically informed in the sense outlined above, but, as a sociologist, I am trying to understand why AI should attract such a diverse range of critics. In the case of gravitational wave detection, for example, what we find are narrow criticisms, sufficiently technically informed to be aimed at the technical claims of the discoverers. Perhaps it is something to do with the human sciences, to which ambitious AI is a contributor: if you can crack the problem of how humans think, it is tempting to imagine you have cracked the problem of everything that concerns humans. AI invites a broad range of types of criticism by presenting itself as all manner of different things, some of which are very grand. Thus, though AI most often presents itself as a narrow technical activity aiming for better engineered devices, sometimes it presents itself as Dawkins-like, anti-religious movement, 'we must show, or have shown, that humans are just meat machines'; sometimes we are told, 'if programs like 'Watson' or 'AlphaZero' can teach themselves to become world champions at any game you put in front of them then our firm's products are what you should be buying' whatever problem you want solved; and sometimes we are told 'the machines are becoming so intelligent that soon we'll be lucky if they are willing to keep us a pets'. All these ways of presenting AI apart from the first are directed primarily at outside audiences and, the first one aside, they don't rest on any deep analysis of what AI has actually achieved in respect of the claim that has been presented.

Note that at the technical heart of the discipline a few experts are willing to admit that no-one yet knows how to make a machine that can reliably manage something as simple as translating, 'The trophy would not fit the suitcase because it was too small', into French;

unfortunately, at the same time, certain insiders claim over and over that the Turing Test has been passed, while those who point out these simple failings don't get anything like the same publicity![4] It is no surprise that the hype has resulted in a series of AI summers and winters and there is growing fear of another cold season on the horizon.

Given this context, the bulk of what I will do will be to try to refine the choice of target for critics and suggest that the AI community would be better off, from the point of view of science, if they presented the aims and achievements of AI in a more precise way. I'll explain my own criticisms of the science of AI, say which aspect of AI they address, and show what it would mean for me to turn out to be wrong. I'll set this technical part of the paper in the framework of a 6-level analysis of what success in artificial intelligence means.

---

[4] The 'it' in the sentence must take feminine gender since it refers to la valise; if the final word had been 'big' instead of 'small' then the gender would be masculine. For Geoffrey Hinton admitting to this kind of problem, but also claiming that it would be solved sooner or later with current techniques see https://www.youtube.com/watch?v=zl99IZvW7rE (around 8.5 minutes in). Kevin Warwick is an exemplar of the Turing Test nonsense and one can rely on the claim being trotted out again after each iteration of the annual Loebner Prize competition. For a detailed analysis of 'Winograd schemas' – of which the trophy/suitcase problem is an example – and why this and other such problems exist, see Collins, 2018, especially Ch 10 and for an account of a recent test turning on Winograd schemas see Levesque, et al, 2012. Ernest Davis (private communication, 14 Sept 2020), points out that been a recent breakthrough in respect of Winograd schemas, machines now coping with around 90% of them; I'll return to this I the Postscript but just mention here that it makes no differences to the arguments in this paper since there are many other similar problems that the machines cannot cope with; a discussion can be found in Collins, 2018 Chapter 10; the 6-level analysis of AI (see below) is also first found in that book.

The big picture

I have to admit, however, that I also have additional audience-directed goals in mind. The first of these is that I don't want people to be so fooled by the hype that they *do* become slaves, not to incredibly intelligent computers but to stupid computers that are taken to be unquestionable authorities – 'the computer says no', syndrome. That's one reason it's important to point out that current AI has not yet reached Stage 3 of the 6-level scale, and that the current debates are all about the transition from level 2 to level 3, with lots of people not understanding how huge even this transition is.

The second audience-directed goal is itself rather grand. I believe the survival of democracy as we have known it in 'The West' depends on the survival of science. This is not the science of Stephen Hawking's *Brief History of Time*, nor even the science of Newton and Einstein, it is the uncertain science responsible for technological decision-making in the public domain, which is not glamorous or even specially likely to be right, but which provides a role model for how to make technological decisions under uncertainty; this is through 'craftwork with integrity'. In these circumstances it is the value system of science and the aspirations of science that are crucial: when we do not have an exact way to calculate the best way forward we need to know that those reaching for it are skilfully trying their best knowing that their efforts will be vitiated by anything

other than the utmost integrity. That is how science proceeds when it is not distorted by the values of other institutions and that is why science is an object lesson for democracy.[5]

The other reason we must support science if democracy is to survive is that, like many other institutions, it is under threat from populism. Elected populists declare they have the mandate of 'The Will of the People' and that the 'checks and balances' that support pluralist democracies, and give some rights to the views of those who were not victorious at the election, can only be traitorous, since they hinder the execution of 'The Will of the People'. Science is one of those checks and balances. This is obvious at the time of writing, with, just to give the most obvious example, President Trump's attacks on environmental and pandemic science – scientific experts limit his freedom to interpret 'The Will of the People' in his own way.[6]

Because of the way it is supported by Silicon Valley, AI is now one of the most secure sciences in the world in terms of research funding. Deep learning is a magnificent success and has conquered some of what were once thought to be high peaks of human intelligence, and one can see it reaching the foothills of what are the true peaks.[7] For this reason, and its location at the centre of technological commerce, AI is a science that is in

---

[5] This argument is made explicitly in Collins and Evans's, 2017 book entitled *Why Democracies Need Science*.

[6] See Collins et al's 2019 book, *Experts and the Will of the People*.

[7] Bert Dreyfus pointed out that AI enthusiasts had a tendency to think that climbing a tree was the first step to reaching the Moon. I am inclined to say that with deep learning we have taken a low altitude flight in a rocket.

the public eye and is going to be more and more in the public eye. Therefore, in so far as science is a role model for democracy and a check and balance, AI can play a more important indirect political role in our lives than most other sciences. AI no longer needs the hype that the orphan enterprise felt impelled to embrace in the early days of the *Lighthill Report* and the like. It is time to make AI a science like physics in terms of the norms of internal criticism sketched out above. AI should become its own most severe critic and set an example for science and for integrity in the creation of knowledge. The huge successes, and the still more huge successes to come in the near and medium future, are not the result of attaining the summits of human intelligence but struggling into the foothills, and it is the science of AI itself that should be telling us this. Now we go back to the nature of human knowledge to see why it is that we are still in the foothills while the summit is a long, long, climb away.

The six levels of AI

The need to separate the aims of AI into six levels arises from the fact that human knowledge is collective. Most ambitious AI aims to reproduce the human brain but without noticing that the brain gives rise to human-like intelligence only when it works within societies of humans; feral children do not develop normal human capacities. For good reasons, the Turing Test is about linguistic ability; language is a collective accomplishment so a properly designed Turing Test will be looking for the ability of the computer to embed itself in human society in the same way that human language speakers embed themselves in order to acquire fluency. That's why deep learning is so good – because it can strip meaning from the continually changing language of the internet (including its latent racism and sexism), and that's also why it cannot manage to

translate that sentence about trophies and suitcases reliably (or similar test) – because it doesn't know the social world of trophies and suitcases and, at the time that example was invented, that social world was not represented on the internet.[8] Given this, one can see that one ambitious aim of AI might be to reproduce a brain that can grasp the collective understandings of any human group into which it is conversationally embedded while a still more ambitious aim might be to reproduce a whole group of human-like AIs, with its own creatively developed collective understandings, not necessarily familiar to existing humans: that's two of the six 'levels' of AI – levels III and level V – shown in Table 1 and set out in more detail below.

LEVEL		PASS TURING TEST (TT)?
I	Engineered Intelligence	We tend not to ask the question
II	Asymmetrical prostheses	Pass non-demanding TT? (Think of Eliza!)
AI IS CURRENTLY TRYING TO MOVE FROM ABOVE THIS LINE TO BELOW THIS LINE		
III	Symmetrical culture-consumers	Pass demanding TT
IV	Humanity-challenging culture-consumers	Pass demanding TT
V	Autonomous human-like society	Pass demanding TT
V1	Autonomous alien society	We would not know how to run a relevant Turing Test

Table 1: Six levels of artificial intelligence (Simplified version of Table 4 in Collins, 2018)

[8] The sexism and racism of computers that strip language from the internet is not a failure but a triumph for deep learning: it shows it has a capacity for socialisation even if what is being acquired is a sad reflection of the embedding society. And, remember, it is easy to make a computer cope with any specific example once the specific problem has been pointed out and even in the act of describing the trophy/suitcase problem the potency of the internet in respect of that example is being enhanced. But the trophy/suitcase problem stands for an indefinite number of new examples which can be invented or occur naturally as society and language changes.

Level I: Engineered intelligence

The first level is the engineered intelligence which we already live with. Bear in mind that some people think that a simple thermostat is intelligent.[9] Engineered intelligences control not only washing machines but power stations and missile launchers, so they have the potential to destroy us by accident without using much in the way of intelligence at all; that's not a criticism of AI, it is just a way of distinguishing the potential danger of computers from their intelligence. Mostly engineered intelligence is a good, life-enhancing, thing that we would hate to be without now we have such a lot of it (I depend on it enormously as I write this article).

Level II: Asymmetrical prostheses

A 'prosthesis', as the term is being used here in the context of AI, is something that fits into society and does the job that a human once did. We might take a calculator as an example – it does arithmetic instead of a human – but so are many of the examples that populate the previous level. So, the only difference between this level and the previous one is the extent to which people *think* that what is going on is real human-like intelligence; what humans think is rarely a matter of sharp distinctions. We can see that a thermostat is at the lower end of Level I and we can see that something like Siri, or Alexa, is at the top end of Level II, and where Level I ends and Level II begins is not that

---

[9] Russell and Norvig, 2003, pps 48-52

important. (In contrast, as we will see shortly, where Level II ends and Level III begins is hugely important.)

Even though the boundary between the first two levels is fuzzy, it is useful to have a Level II because it makes what is meant by 'asymmetrical' a little clearer. A crucial feature of human interaction with other humans, and with machines and other material objects, is what is called 'repair'. Starting with humans, when I am muttering indistinctly to you, you will mostly manage to work out what I am saying from the context without having to ask me to clarify. That's repair: I speak in some kind of broken or incomplete way and you use the context to fill in the gaps and file off the sharp edges to make a smooth and well-formed piece of communication. Without that, every act of communication would have to be perfect, or we would have to be continually repeating ourselves, and this would make communication very cumbersome.

This human talent for context sensitive repair is also continually deployed in our interaction with machines. For example, when we think, 'my calculator is a lot better at arithmetic than I am', it is because we continually repair its mistakes without noticing. To give an example from my 1990 book, if I want to know my height in centimeters, given that my height in inches is 69, and there are 2.54 centimeters to the inch, and I key 69*2.54 into the calculator, it returns, '175.26' in an instant – better than I could do in an instant – but 175.26 is not my height (at least, not for more than a fraction of a second between breaths and depending on the state of my hair), but I unthinkingly repair it to 175cms. So in that sense the calculator is not as good at arithmetic as me because it does not know how to understand social context in a way that would cause it to approximate

appropriately in the context of discussion of human height – part of the skill of good arithmetic. The calculator does not understand social context of arithmetical calculations of human tallness in the same way as deep learning translators do not understand the social world of trophies and suitcases.

Now, this is important, because AI 'boosters', not to mention various misguided philosophers, psychologists and sociologists, think that anything that enters our social life and has an effect on it as a prosthesis – and calculators, word processors, Siri and Alexa, certainly do have such an effect – should be treated as social creatures. They want them to be treated as nodes in the networks of relations that describe our social lives that are indistinguishable from other humans or, even treated as full-scale social intelligences.[10] But they are not full-scale intelligences because if you *talk to them* in context-dependent and otherwise damaged ways they won't be able to repair *your* output in a satisfactory way. Predictive text and spell-checkers do their best at repair, indicating that their developers know there is a problem, but they are clunky toys rather than serious contenders.

Level III: Symmetrical culture-consumers

So that is why it is useful to have a Level II category of asymmetrical prostheses even though it is not clearly distinct from Level I. Level II, to repeat, turns on prostheses, the

---

[10] For those who know the field of STS (Science and Technology Studies), Latour's so-called 'actor network theory' is the most popular and notorious example of this kind of elision but it is a very widespread mistake even among those who do not choose to elevate it to the central plank of a theory of the world.

output of which we find effortlessly useful because we automatically repair that output without noticing it, just as we do with other humans. It is useful so that we can contrast it with the category of *symmetrical* prosthesis, which can effortlessly repair *our* output with as much context-sensitivity as we can repair theirs; this is Level III. At the moment, Level II AI is continually being confused with Level III and trumpeted as being real artificial intelligence, passing the Turing Test, and so on. But the jump from Level II to Level III is huge and so is the jump from where we are to serious artificial intelligence: as can be seen, it is going to involve the AI's being effortlessly embedded into society so they can understand social context as well as humans understand social context; when the can do that, they will be able to pass properly designed, demanding, Turing Tests, rather than demonstrate a facility with games. When they can do that, they will be able to absorb human culture in the way that humans do. That is why, in Table 1, they have been given the label, 'symmetrical *culture-consumers*', rather than prostheses.

Impact of AI on our understanding of knowledge

In the Introduction I said I thought that the science of AI had already taught us a lot about knowledge and intelligence. So much has it taught us that I think the philosophy and psychology of skill and expertise must change to take account of it; we now have to talk of 'knowledge' not '*human* knowledge' if we want to understand any kind of knowledge including human knowledge. Before AI came along, the philosophy of skill and expertise had to do only with what humans could do; to study skill and expertise was to study humans. But now we need to change the focus away from humans and to the knowledge stuff itself. To understand knowledge, we must understand what machines can and cannot do just as much as what humans can and cannot do. For example, one

huge change wrought by AI research, albeit inadvertently, is to our understanding of what counts as the apogee of knowledge. Before AI the apogee was taken to be somewhere around the things that humans find really difficult and high accomplishment in which was lauded and rewarded: when I was a school, this was the ability to do mental arithmetic and in adults it was the ability to do a really tough integration or some other such mathematical *tour de force*. But now we can do that kind of arithmetic with a pocket-sized calculator and the program *Mathematica*, can manage the fancier stuff, so that kind of thing is no longer seen as the apogee. The apogee is now seen as somewhere around some of the things that humans find easy and are still beyond computers – such as fluent language speaking.[11] Deep learning's huge success in improving language handling is therefore very impressive but its failure to handle such things as trophies and suitcases shows how far there is to go: we are still at Level II rather than to Level III even though the Level II accomplishments are nuzzling the boundary.

A good way to see the difference is to consider the accomplishments of AlphaZero, which taught itself to be world champion at Go and at Chess in a matter of days, again, accomplishments that were thought to be the apogee of human accomplishments until very recently and may still not have fallen in most peoples' estimations given the short time that has elapsed since these peaks were conquered. But both Chess and Go, even though the perfect game cannot be calculated through to the end, in the way that Noughts

---

[11] Except in the films where all the intelligent computers and robots are effortlessly fluent language understanders and speakers even while they are psychopaths: Hector Levesque's (2017) definition of AI, 'the study of how to make computers behave the way they do in the movies', has it about right.

and Crosses can, are still played in a fixed format according to fixed rules with a fixed end-point. To reveal Level III ability, when you sit down to play Chess or Go against AlphaZero in April 2020, it would have to do what humans do and make a bit of small talk about how things were going for you in the Coronavirus crisis, before it even thought of making a move. It would have to know what fluent social interaction comprised in the current context. AlphaZero is still somewhere in the first two Levels.[12]

Who should be criticising AI?

The complaint from AI enthusiasts will be along the lines that 'every time we accomplish something new and magnificent, we're told by the critics that if it can be accomplished it can't be the real thing after all.' And something has gone wrong if the critics are doing this continually – it they are continually making the goal of human-like AI an ever-receding target. But the thing that has gone wrong isn't to be found in the critics' domain; the thing that has gone wrong is that it should be the AI enthusiasts who are getting in first and pointing out what they have not yet accomplished in spite of the fancy and unexpected results, even when those results reach a target that the critics said could not be reached. That's how other sciences go: in those sciences the aim is exactly defined and the worst sin is to make the claim that the aim has been accomplished and

[12] As Colin Garvey pointed out to me, there is a well know critical remark from the 1970s to the effect that an ideal AI of the time could make a perfect chess move while the room was on fire. There are many examples of what Level III AI's need to be able to do in the way of editing text but Level II AI's can't do in Chapter 10 of *Artifictional Intelligence*.

have it turn out not to be so.[13] Maybe, as some of them claim, the deep learning community will get there through a huge increase in artificial brain capacity, maybe not.

Level III to Level IV

Notice that the aim of AI in which I am interested is learning to understand human-like intelligence by trying to simulate it with non-human means. The criterion that I am taking to indicate the achievement of human-like intelligence is passing *demanding* versions of the Turing Test – 'DTT's that demonstrate a grasp of social context and the corresponding ability to repair broken speech in a human-like way. Now, it seems to me that, currently, the most promising route to this goal will include the building of machines that mimic the mechanisms of the human brain in some abstract sense; I am impressed by the argument that what we need to build are better and better versions of hierarchical pattern recognisers. But that's not where my expertise lies so my hunch in this respect is not worth much. It may be that the internal mechanism of the artificial entity that meets my criterion will be different to that of the human brain. Ava, the AI imagined in the film, *Ex Machina*, Samantha, from the film *Her*, and HAL, as portrayed in *2001, A Space Odyssey*, are thoroughly context-sensitive, fluent English speakers (who just happen to be

---

[13] And to be crystal clear, when AI reaches Level III, and according to at least some people who do understand what a demanding Turing Test would look like, they think this will be accomplished pretty soon, perhaps even in my lifetime if I am lucky, I will be delighted and ready to say that human intelligence has been simulated artificially. (What a demanding Turing Test would look like is explained in Chapter 10 of my 2018 book.) That is, I will be delighted provided the AI enthusiasts have not surreptitiously shifted the goalposts themselves by aiming to convince an audience with some tricks rather than truly aiming for linguistic fluency in the face of the hardest, context dependent, tests. Sadly, this kind of goalpost shifting also happens all the time.

psychopaths). In at least two of the cases, Samantha and HAL, they are imagined as being dissimilar to humans in terms of their physical construction. Nevertheless, they are still Level III devices and meet my criterion of mimicking human-like intelligence. Such devices might be a little disappointing in that we may not learn as much about human intelligence from mimicking it in this way as we would have learned from mimicking the human mechanism, but maybe in recognizing we are doing something different we will still be learning what the real mechanism is.[14] The point is, that there is also a Level IV where the mechanism is broadly the same.

Level IV of Artificial Intelligence: Humanity-challenging culture-consumers

The difference between level III and level IV is a subtle one and hard to pin down. At Level III AI achieves human-like intelligence but at Level IV the mechanism by which it achieves it must be the same as when it is exhibited in by humans. The question of similarity of process will remain important for those interested in AI as a route to proving that humans are just meat machines without free will, and that humanity doesn't have a soul nor anything unique that could stand in for it. It will also remain important for those who think AI is the route to understanding human intelligence even if they are not pre-committed to any metaphysical view.

---

[14] The idea of interactional expertise is important here. Some argue that a human-like body is necessary to achieve human-like understanding sufficient to pass a Turing Test. The question is discussed in, for example, Collins,2020, but this argument probably has some way to go.

Unfortunately, whenever the notion of 'the same' is deployed there is always going to be a problem about what it means. For example, would a Level IV device need human-like consciousness (which is not a precondition for Level III)? Since humans themselves can carry out the same actions with varying degrees of consciousness (eg it is claimed that a mark of truly skilled physical performance is an absence of conscious attention), it is hard to foresee where the argument about the need for consciousness will go in the case of Level IV. If we are not sure if consciousness is a precondition for intelligent action in humans how can we claim it is a precondition for achieving intelligence in the same way as humans achieve it?

One must not make the answer to what constitutes the human process a truism by insisting that doing things like humans means using the same *biological* mechanisms or the problem becomes not one of reproducing *human intelligence* but reproducing *humans*. So, we must accept that thinking 'like' a human while using silicon chips or some such, potentially meets the criterion of Level IV – reproducing human internal states. But what about AlphaZero?; setting aside the small-talk problem, you could not play a decent game of Chess or Go with it because it would win every time. It could be claimed that it is still using the human thought processes that humans use when they play games – hierarchical pattern recognition – but just doing it much better! So, nice philosophical questions remain at this transition point.

Levels V and VI

Level V is like Level III, or Level IV, except that the AI's will be sufficiently human-like so that groups of them can develop human-like cultures by

themselves. Here the question of the necessity of a human-like body, which is disputed at Levels III and IV, is resolved. For a group of machines to develop human-like cultures they will need human-like bodies because, while it can be argued individuals can *acquire* human-like culture through immersion in language alone, without participating in the physical activities of a culture (interactional expertise), we can't *develop* human-like cultures, nor sustain them over the generations; we must have a body-type which affords the corresponding physical activities.[15] So such machines will need bodies that could play tennis and cricket and American football and snooker, and so on, at about the same level as humans, even if they don't wind up playing them but develop their own sports or reject sport altogether. Solutions to the problems of human-like robotics are going to be essential at Level V. Thus, while an autonomous society of intelligent dogs – dogs with a more elaborate speaking apparatus – might develop a new language, it couldn't include words for tennis racket or cricket bat, at least, not sustainably, unless the dogs encountered humans and maintained linguistic contact with them.

Humans continually try to develop new cultures. Sometimes these turn on physical appearance. Thus, there is currently a half -jokey movement under way to bring out the cultural specificity of red-haired people. So, let us equip our otherwise human-like AIs with a metallic-looking silver skin so they can be

---

[15] For interactional expertise see Collins and Evans 2015 and Collins, 2020

easily identified by others and easily recognise each other. We can then imagine them forming their own cultural group, proudly or even aggressively distinct from the other human cultures around them and pulling away from existing human societies (this, of course, is also the stuff of science fiction). That would be Level V of artificial intelligence.

This seems to be what those fearful of the 'singularity' are thinking of when they claim that the computers will one day be 'so intelligent' that we will be lucky if they are willing to keep us as pets. The doom-mongers see intelligence as a monotonic accomplishment, of which you simply have either more or less, and if you have more you automatically become more powerful and dangerous. But there has to be something special about the intelligence if it is going to be inclined to overpower the humans who made it; it has to be the kind of intelligence that is capable of forming its own cultures. That culture may not be a violent one; perhaps it will be a peaceful culture – there are many such. But it may pick up its cues from the violent intentions in human societies. It is probably sensible not to try to make Level V just in case it does develop in a way we will regret.

Level VI departs from the AI aspiration set out at the beginning of the piece "attempt to explicate the workings of human intelligence by creating 'mechanical causes and effects that mimic human actions'"; Level VI may still be trying to explicate the workings of human intelligence but, if it is, it will be doing it by creating non-human kinds of intelligence. Taking our cue from science fiction, once more, it might try to

create an artificial version of the intelligence of the extra-terrestrial heptapods portrayed in the film *Arrival*. It will be difficult to know whether we have created such an intelligence because we would not know what questions to ask in a demanding Turing Test; it would be like a Turing Test to distinguish between a machine mimicking a Chinese-speaker and a human Chinese-speaker but where the judge does not speak Chinese. I don't know whether Level VI really is within AI's project, but sciences do develop their own momentum and it is not hard to imagine AI going this way once it has reached the other levels. Once more, it could be a hazardous undertaking.

Conclusion

What I have tried to do here is point out some features of AI and its discontents and compare them with the critical debates in other kinds of science. Compared to say, physics, the debate about AI is diffuse. I have suggested that this is because both proponents and critics of AI most often direct their arguments at outside audiences rather than inside practitioners. A stark contrast is found in the fact that in domains like high energy physics, or gravitational wave physics, it is recognised that the sharpest and most productive criticisms have to come from inside the domain whereas all too often in the world of AI insiders, hyping of products takes precedence over internally organised criticism. This leaves the field free for outsiders to generate a heterodox range of complaints, and it leaves the field vulnerable to these complaints. There is a tradition of insider criticism of AI, and, unsurprisingly, it is sharp, but the tradition seems to be getting thinner.

My criticism is based on my expertise on the nature of human understanding – that which is to be mimicked by AI – and I use it as a way of refining the aims of AI and dividing them into a possible six levels. My claim is that we are currently at the top end of Level II but still a long way from Level III. Attainment of Level III will be demonstrated by AIs that can pass suitably demanding Turing Tests (DTTs), which currently no machines can pass. The continual claims by AI boosters that the Turing Test has been passed is a problem for AI as a respectable science and as the `Western World' encounters political dangers that most of us thought had long become part of history, we desperately need sciences to act respectably so they can be role models for decision-making and legitimate checks and balances or political ambitions . The six levels may function as a way of bringing some order to the ambitions and claims about the accomplishments of AI. No doubt other ways of defining the aims of AI could be devised, though I think all of them will need to include the division between Level II and Level III. Whichever way of defining the aims gains the most widespread assent, to have a more carefully defined target should help, and that could be vitally important for far more than AI itself.

In sum, AI is one of the most important sciences in the world but its way of presenting itself is still too influenced by its early insecurities as a science. To play the role we need it to play in today's political world, a role which it can now well afford to play given its almost unprecedented financial independence as a science, it needs to curb the reflexes developed in those early days. Instead it should act like the iconic sciences of physics: that is, the aim should be never to announce that more has been achieved than has been achieved. This aim can be achieved only through the nurturing and honouring of internal critics rather than the rejection which was typical of the formative period. I have to add

that it should be obvious by now that human intelligence is a collective enterprise with language being an iconic example: those internal critics will have to take this into account, emulating brains that don't stand alone but interact fully in society. If AI can switch to becoming this kind of strongly self-critical science, future generations would look back at it as one of the institutions that helped save pluralist democracies from populism rather than as a notorious champion of alternative facts.

Postscript: Marcus and Davis; engineering and the social.

After the penultimate draft of this piece was written I came across the recently published book by Gary Marcus and Ernie Davis entitled *Rebooting AI: Building artificial intelligence we can trust*, and also their critique, in *MIT Technology Review,* of a recent and much-hyped language processing device known as GPT-3 (GPT = stands for Generative Pre-Training). Since both Marcus and Davis (M+D) are technical insiders in the field of AI and their work is interestingly and productively critical, and since it illustrates some of what is argued above, it seemed useful to discuss it. I referred to some of Davis's fascinating 'common-sense' criticisms and his decisive 'Winograd schema challenge test' for existing AIs in my 2018 book (*Artifictional Intelligence* – hereafter '*AA*') on which my article is based). Their book is also reassuring in that these technical experts include many of the same technical elements in their discussion of AI as can be found in *AA* though, of course, they are able to include far more detail concerning technical developments.

Starting with GPT-3, it is striking that the company that makes it refused to allow M+D access to it for the purpose of testing it; they had to obtain access to it through a 'back door'. Here we see, once more, AI still not acting as a respectable science encouraging technical criticism. As M+D explain in their 2019 book,

> Silicon Valley entrepreneurs often aspire to "move fast and break things"; the mantra is "Get a working product on the market before someone beats you to it; and then worry about problems later." (p188)

Once M+D were able to test GPT-3 it on questions of common sense and the like, they were able to show that it represented no significant improvement in terms of language 'understanding' over previous language processors in spite of the hype: it's failures were of the trivial type discussed above (and in Chapter 10 of *AA*).[16]

But their book is also revealing in the ways that it differs from mine and from the light it sheds on some of the arguments presented above. The differences arise, I believe, out of disciplinary background and approach: Marcus is a psychologist by training and Davis is a computer scientist. Furthermore, both approach the topic as an engineering problem. They state at the outset of their book:

> Crucially, AI is not magic, but rather just a set of engineering techniques and algorithms, each with its own strengths and weaknesses, suitable for some problems but not for others (p24)

[16] Though, GPT–3 is much better than previous intelligent machines at handling the narrow class of demanding Turing Tests that turn on Winograd Challenges in particular (thanks to Ernie Davis for pointing this out to me). Nevertheless, there are a number of other 'simple' language challenges that Marcus and Davis point out are still beyond GPT-3 and even the success re Winograd Schemas seems to be a matter of ad hoc engineering solutions – Band-aids – rather than a deep breakthrough; thus it takes success to the 90% level whereas humans' success is normally around 100%; this seems a difference in quality not quantity. There is, however, a nice argument to be had about whether the steady achievement of success of this type is actually demonstrating how humans think but, as pointed out in the paper, it is hard to generate real engagement.

In contrast, *AA* approaches the topic from the point of view of a philosophically inclined sociologist and looks at AI as a science aimed at elucidating the nature of human knowledge.

Starting with the psychology/sociology tension, though the topic of the relationship between top-down and bottom-up understanding appears in both my book and theirs, they present it as a matter of individual psychology, nicely demonstrated by the way the same individuals can be 'primed' with different stimuli to interpret an image in different ways. There is no discussion in their book of how whole societies, or sub-groups within societies, or adherents to different scientific paradigms, inhabit different social settings in which the world is viewed in the same way by all the inhabitants of that setting but in different ways to inhabitants of other social settings. This individual/collective contrast is a key to the difference in thinking.

Turning to the engineering versus philosophy/sociology contrast, M+D's aim is to improve AI by preventing its being taken over by deep learning techniques which start every new task from scratch. Instead, they believe successful programs must use older AI techniques to insert a large component of explicit physical and common-sense understandings as a foundation for any subsequent learning. Here we can see the influence of Marcus's mentor, Steven Pinker and, in turn, his debt to Chomsky. These older techniques have not proved effective on their own but, they argue, should be much more effective when *combined* with deep learning. They also suggest inserting, 'by hand', some ethical principles into programs from the start. They are engaged, then, in

an internal technical battle to stop deep learning entirely taking over the world of AI and countering this tendency with the addition of some more explicit programming.

In *AA* (p110 ff) I argue that to make sense of the different way different groups of humans interpret the world there must be a common foundation of perceptual abilities on which the varied interpretations are based, so I sympathise with M+D's desire to base deep learning on an explicit and universal perceptual foundation. But the big question is how deep this foundation should be.

M+D appear to want to build a deep foundation based on the knowledge of how Western societies work and including both lots of scientific understanding as well as norms of behaviour in different settings. This approach might well produce better engineered AIs for use in Western settings but the resulting programs are still going to be vulnerable to occasional, unpredictable, unhuman-like failures whenever they approach tasks in unrestricted or new domains and settings. Ironically, M+D are themselves experts on these kinds of problems as Davis's excellent Turing Test challenges quoted in their book on page 93, and in Chapter 10 of *AA*, and which forms the basis of their critique of GPT-3, reveals. They know the human world is open-ended, that the knowledge and common-sense that forms it cannot be captured in a set of formal rules and that, therefore, the addition of sets of explicit rules and facts may improve intelligent devices but only in the way that a Band-aid improves a wound; they know this as well, or better, than anyone else. And yet they seem to forget this issue when they offer advice for making a better AI. The schizophrenia is also there when they insist that a replacement for the Turing

Test is needed given that Davis's own versions of a demanding Turing test is the very tool they use to show the deep inadequacy of existing AIs.[17]

It seems to me that the schizophrenia arises from approaching the problem as one of engineering rather than philosophy/sociology. That goal leads them to try to work out a way to build programs that will better capture features of Western scientific culture. What they fail to notice is that different groups of humans see the world in very different ways; one cannot fail to notice this if the goal is to reproduce human intelligence. An AI that is to reproduce human intelligence will have to be itself *capable* of seeing the world in many different ways. Building AIs that start with the uniform model of the world provided by the current state of Western scientific culture will not solve the problem of human intelligence. Any built in foundation of common perceptual abilities, such as the recognise basic shapes and patterns and so forth has to be a shallow if it is to allow scope for all the varied perspectives of current and future human groups, if Ais, like humans, are to be able to learn different things from exposure to different social worlds.

On page 201 of their 2019 book, M+D state:

---

[17] I have argued at length (eg in *AA*) that a computer that can pass a demanding Turing Test (rather than some tricksy version of it), will have solved the deep problems of AI. This depends on understanding the idea of 'interactional expertise', which argues that an understanding of human practical abilities, if not the ability to practice, are captured in language native languages (including technical languages). This makes a demonstration of fluency in a language a demonstration of what counts as intelligence. See, eg., Collins and Evans, 2015; Collins, 2020 for 'interactional expertise.'

> AI that is powered by deep understanding will be the first AI that can learn the way a child does, easily, powerfully, constantly expanding its knowledge of the world (p201)

How can one disagree that 'deep understanding' is a likely component of human-like intelligence? How can one disagree that learning 'the way a child does, easily, powerfully, constantly expanding its knowledge of the world' would be a fine thing in an intelligent computer? But this is like agreeing with the virtues of motherhood and apple pie. What does 'deep understanding' mean and why is there no discussion in their book until page 201 of how *humans* come by their common-sense? Why is there no discussion of how you would build a computer that could occupy the social spaces that children occupy in the course of their upbringing so that they could truly learn like a child? Why is there no discussion of how this will lead to the marked variations in the substance of the intelligences of such devices when immersed in different social locations? I am suggesting that it comes from mixing up engineering solutions with understanding human knowledge. In open domains, engineering solutions are always going to be a matter of more and more Band-aids.

In sum, to get beyond Level 2 of artificial intelligence, and find the kind of solutions that the non-engineering side of Marcus and Davis want, which would pass their demanding Turing Tests, will require building machines that are capable, in principle, of absorbing non-Western cultures as readily as Western cultures and have the potential to absorb all the varied, cultures (including the crazy ones), found in Western societies. If a machine is to do that, it cannot be constrained by too much built-in current science and

engineering so its knowledge and rules foundation will have to be a shallow one and the large preponderance of what it knows will be learned from scratch and therefore capable of being different in different settings.

When a machine has been built that can absorb all these different cultures – which means a machine has been built that can truly learn like a child – then it will be also be a machine that can absorb Western scientific culture properly rather be pre-programmed with a stick-figure caricature of Western scientific culture. It will then be able to handle the engineering problems presented by Western cultures as reliably and creatively as humans while its failures will be human-like. Ironically, then, really good engineering solutions will need to solve the philosophical and sociological problems first!

References Cited

Collins, Harry, 2020, Interactional Imogen: Language, Practice and the Body, *Phenomenology and the Cognitive Sciences*, 00, 00, 0000-0000, DOI: 10.1007/s11097-020-09679-x

Collins, Harry (2018) *Artifictional Intelligence: Against Humanity's Surrender to Computers*, Cambridge: Polity Press

Collins, Harry (2017) *Gravity's Kiss: The Detection of Gravitational Waves*, Cambridge Mass.: MIT Press

Collins, Harry, (2010), *Tacit and Explicit Knowledge*, Chicago: University of Chicago Press

Collins, Harry and Evans Robert (2017) *Why Democracies needs Science*, Cambridge: Polity Press

Collins, Harry and Evans, Robert, (2015), 'Expertise Revisited I - Interactional expertise' *Studies in History and Philosophy of Science,* 54, 113-123 (a pre-print is available at http://arxiv.org/abs/1611.04423)

Collins, Harry, Evans, Robert, Durant, Darrin and Weinel, Martin, (2019) *Experts and the Will of the People: Society, Populism and Science*. Basildon: Palgrave

Levesque, Hector 2017. *Common Sense, the Turing Test, and the Quest for Real AI*. Cambridge, MA.: MIT Press.

Levesque, Hector, Davis, Ernest and Morgenstern, Leora 2012. *The Winograd Schema Challenge*. Proceedings of Principles of Knowledge Representation and Reasoning.

Marcus, Gary and Davis, Ernest 2019. *Rebooting AI: Building artificial intelligence we can trust*, New York: Vintage

Marcus, Gary and Davis, Ernest 2019. 'GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about', *MIT Technology Review*, August 22, 2020

Russell, Stuart J. and Norvig, Peter 2003. *Artificial Intelligence: A Modern Approach* (2nd edn). Upper Saddle River, New Jersey: Prentice Hall.