ANNALES

PROCEEDINGS OF THE ACADEMY OF SCIENCES OF BOLOGNA

CLASS OF PHYSICAL SCIENCES

Bologna
University Press

ANNALES

PROCEEDINGS OF THE ACADEMY OF SCIENCES OF BOLOGNA

CLASS OF PHYSICAL SCIENCES

1

Annales. Proceedings of the Academy of Sciences of Bologna Class of Physical Sciences

Fondazione Bologna University Press
Via Saragozza 10, 40123 Bologna
tel. (+39) 051 232 882 – fax (+39) 051 221 019
ISSN: 2975-2302
ISBN: 979-12-5477-324-6
ISBN online: 979-12-5477-325-3
DOI: 10.30682/annalesps2301

www.buponline.com
info@buponline.com

Cover: Pellegrino Tibaldi, *Odysseus and Ino-Leocothea*, 1550-1551,
detail (Bologna, Academy of Sciences)

Layout: Gianluca Bollina-DoppioClickArt (Bologna)

First edition: December 2023

Table of contents

Stream-of-consciousness thoughts on language and AI

Marina Frasca-Spada

Corpus Christi College, Cambridge

Contribution presented by Pierluigi Contucci

Abstract

I consider the shifts in the meaning of some of the terms in use in discussion of AI-related themes at different levels of technical information, from the levels of specialists to those of the general public. I identify these shifts as one cause for the confusion and contradictions dominating such debates. I conclude highlighting the need – and the opportunity – for a renewed closer cooperation between scientists and philosophers.

Keywords

Artificial intelligence, Intelligence, Discovery and creativity, Language, Social machines.

1. A thermostat's beliefs

It is not unusual for a new and very dynamic discipline to call for a renewal of philosophical language to accompany the creation of its new conceptual toolbox and corresponding specific terminology (no need to invoke Thomas Kuhn or Ian Hacking here). But then it is perhaps unsurprising that current discussions on AI and allied disciplines' philosophical, ethical, sociological import and consequences involve much negotiating the exact meaning of key terms, most of them borrowed from other disciplines or ordinary language. Such are, for instance, "intelligence", "creativity", "intuition" and "discovery", "causation", as well as "rights" and "ethics", "belief", "technology" and more, all of whose meanings are being variously discussed, questioned, or even off-handedly replaced by more appropriate, but less generally familiar ones.[1]

Consider "belief". I found it defined, in passing, as "model of the environment" (*e.g.* Burr *et al.* 2018, 741). This is entirely sensible, although rather unusual for ordinary English speakers. In fact, a new definition of "belief" was a relatively early occurrence in the creation of an AI-specific technical terminology. In 1979, John McCarthy was explicit about the need to give new definitions of mental concepts and offered a splendidly pragmatic list of (epistemological) reasons why "ascribing mental qualities such as *beliefs*, *intentions* and *wants* to a machine" was both OK and very useful (McCarthy 1979, 15-16); and he went some way in doing just that for "belief" (McCarthy 1979, 12-14). Famously, he wrote that "machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines capable of problem-solving performance". McCarthy cautiously prefaced this somewhat startling point with the qualification that "beliefs" for machines are constructed "in a simpler setting than for humans", and that, although "we will need to build into [a generally intelligent computer program] a general view of what the world is like", "as much as possible, we will ascribe mental qualities separately from each other instead of bundling them in a concept of mind" (McCarthy 1979, 3).[2] For all the intelligent prudence thus displayed, it was in reply to this paper that John Searle wrote his classic and much discussed one on the Chinese room argument (Searle 1980).

Something similar applies to "technology". Normally the term is intended, as one finds in the *Oxford English Dictionary*, as

> **a.** The branch of knowledge dealing with the mechanical arts and applied sciences; the study of this. […]
> **b.** The application of such knowledge for practical purposes, esp. in industry, manufacturing, etc.; the sphere of activity concerned with this; the mechanical arts and applied sciences collectively. […]

[1] This is clearly a very different exercise from that, which Turing condemned as leading to absurdity, of defining the meaning of the terms "think" and "machine" in order to answer the question "can machines think?" (Turing 1950, 433).

[2] In the same mood, McCarthy 2006 spells out with admirable clarity how AI and philosophy can help each other and clarifies that for AI purposes "mind has to be understood a feature at a time" (McCarthy 2006, 2). (David Hume is smiling in his neoclassical grave.)

> **c.** The product of such application; technological knowledge or know-how; a technological process, method, or technique. Also: machinery, equipment, etc., developed from the practical application of scientific and technical knowledge; an example of this. Also in extended use. […][3]

And here’s a (post-Latourian) working definition offered, in passing, in connection with AI:

> […] “technology” does not refer just to an algorithm, but rather to the complex of people, norms, algorithms, data and infrastructure that are required for any of these services to exist. Addressing the current challenges in AI may require adapting all of the above. (Cristianini 2019, 2)[4]

One could easily carry on in this vein. No surprises here.

2. Intelligent action

And now for “intelligence” itself. Discussions on AI highlight how strongly anthropomorphic our usual notion of intelligence is, even among cutting-edge scientists. And it is acknowledged that some serious re-thinking would be timely:

> It is the first time in history that humanity is approaching the challenge to replicate an intelligent and autonomous entity. This compels the scientific community to examine closely the very concept of intelligence – in humans, animals, and of the mechanical – from a cybernetic standpoint. (Veruggio 2006, 612)

For example, what about: “Intelligence as the capacity for autonomous purposeful/ teleological/ goal-oriented/ rational behaviour or action”.

This is a combined version of various definitions to be found in publications relative to AI, all of them fully in line with James Albus’s suggestion, over thirty years ago, that intelligence is

> the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success. (Albus 1991)

Note how “intelligence” is identified without residue with “intelligent behaviour” or “intelligent action”. That this is the focus is unsurprising and, just as in the case of belief, it is evident from the very beginning of AI and across the board. For example, in the classic paper by Newell and Simon we find that

[3] See https://www.oed.com/view/Entry/198469?redirectedFrom=technology#eid (accessed 14 May 2023). (This is no. 4 in the *OED* list of meanings for the term, and the first one that is not obsolete).

[4] As intimated in the text, this has a clear some resonance in Science and Technology Studies and in the Sociology of Scientific Knowledge.

> […] intelligent action […] is, of course, the primary topic of artificial intelligence. […] we measure the intelligence of a system by its ability to achieve stated ends in the face of variations, difficulties, and complexities posed by the task environment. (Newell and Simon 1976, 83)

Before I even start wondering whether I like this new, technical definition of "intelligence", I should declare that there is one particular set of its consequences that seems to make very good sense: *i.e.*, if this is what intelligence is, then we are literally surrounded by non-human forms of it. And it is interesting that both we and they fail to recognise each other as "intelligent": based on the definition above it is suddenly evident that all forms of animal and plant adaptation are forms of intelligences, for all that we routinely fail to recognise them; and it is clear neither do they, for their part, regard what we consider the choicest results of human intelligence as anything in particular, let alone think of them as resulting from intelligent behaviour. Moths do mercilessly and presumably, from their point of view, intelligently make holes in the most extraordinary Kashmiri embroideries, if these are on wool, and if we don't stop them think of what pigeons do on, say, Donatello's statues – either of which, incidentally, may or not be terribly intelligent productions according to the definitions above. Similarly, there is no particular reason to think that if we came across alien extra-terrestrial intelligences we would recognise them as intelligent; and of course why or how would an alien intelligence, no matter how intelligent in the sense above, recognise our sciences or our works of art as productions of a form of intelligence? I could carry on (see *e.g.* Cristianini 2023, 10 ff. for reflection on much of the above; also Darwiche 2018).

And here is a thought that may well seem rather banal, but still I think well worth a passing mention: I find this notion of intelligence, with its welcome emphasis on evolutionary adaptation and hence survival, sobering, because it contains a half-hidden (or perhaps very evident?) element of brutality. Our absent-minded failure to recognise animal and plant intelligence presumably has something to do with the fact that our – human – intelligence has equipped us with an almost unlimited power over them, and the capacity for their destruction (which we exercise all too freely). It is not surprising, then, if the prospect, just round the corner, of an AI that seems completely alien, perhaps not completely under our control, and more powerfully "intelligent" than we can even make sense of (although perhaps in ways that may seem rather muscular and unsubtle), is greeted with a certain apprehension. So as the enthusiasts cite Isaac Asimov's R. Daneel Olivaw and the three laws of robotic governing his positronic brain, many of the worriers seem to have in mind HAL – or, even worse in a *crescendo*, the Westworld Gunslinger and Roy Batty.

3. *Che sì e no nel capo mi tenzona*

Be that as it may, how do we feel about this definition of intelligence: do we like it or not? As mentioned above, I for one find aspects of it very appealing. Even so, I am uncertain. I very much like its down-to-earth-ness, and its inclusivity seems to me both smart and persuasive. But I find it also rather too quantitative – intelligence by the kilo. How does this definition relate to the concept of intelligence underpinning the various quantitative measures of intelligence – IQ tests and the like? Perhaps a bit too closely for comfort.

Also, I have the impression that this is a very pared down conception of intelligence, and one that seems to leave out things that I do think should be included (see above: Kashmiri embroidery, Donatello's statues, presumably Shakespeare's sonnets and the like, or indeed the AI programmes too – as well as most of philosophy…).

On the other hand,

> In AI research one must treat simple cases of phenomena, *e.g.* intentional behavior, because full generality is beyond the state of the art. Many philosophers are inclined to only consider the general phenomenon, but this limits what can be accomplished. I recommend to them the AI approach of doing the simplest cases first. (McCarthy 2002-05, 2-3).

Compare with Galileo's celebrated statement: "quando il filosofo geometra vuol riconoscere in concreto gli effetti dimostrati in astratto, bisogna che difalchi gl'impedimenti de la materia".[5] As well as preaching thus wisely, McCarthy was also phenomenal at implementing his own lesson, one step at a time, one paper at a time, to seriously impressive cumulative effect (although he was still into trying to model human intelligence and now, at least for the moment, we are not into doing that).

4. The quintessential Other

Focusing further on the specificity of AI, this is what the "I" in it stands for:

> Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. (EU HLEG AI 2019, 36)

In this definition AI systems are obviously goal-oriented; they are autonomous in the sense that, although their goals are given, no instructions are issued on how they are to achieve them; and in order to respond to the environment they need to be able to receive input from it – *i.e.*, in some sense, they need to "sense" it (Cristianini *et al.* 2023, 91). In other words, they need a "body". No need to add that in this connection, just as in the case of "sense", so this "sensing" "body" is not to be understood in the usual way as a physical thing, let alone as a particular arrangement of organic matter.[6] Here goes:

[5] Galileo Galilei, *Dialogo sopra i due massimi sistemi del mondo, tolemaico e copernicano*. https://it.wikisource.org/wiki/Pagina:Le_opere_di_Galileo_Galilei_VII.djvu/242.

[6] This is something on which, by contrast, Searle insisted; and presumably, despite the clear disclaimer, he would regard this interpretation of "body" as a disguised form of strong mind-body dualism. See Searle 1980, 423. Also, among others, discussion in Chalmers 2022 (in connection primarily with consciousness/sentience), and Jacquette 1989.

> […] an agent needs a “body”, that is, a way to interact with its environment, and it does not make much sense to consider “disembodied” intelligence (with all due respect for Descartes). However, we do not require either bodies or environments to be physical […] The body is just whatever allows the agent to affect, and be affected by, its environment. (Cristianini 2023, 6)

It is perhaps worth observing that in this view a key difference between AI systems’ intelligences and animal/plant intelligences is that, at least for the time being, only the latter are “embodied” in the same (boringly traditional, hence easily recognisable) way as we are.[7] And yet we – humans –although all but oblivious of “intelligence” in plants and animals, are very ready to be impressed with some AI systems’ intellectual feats and to recognise them as intelligent, indeed we often manifest a strong tendency to think of some of them as actual minds, with consciousness and all.[8] The obvious case is Blake Lemoine with the generative AI system LaMDA; and even the persistently level-headed and sceptical Contucci (2023, 68ff.) acknowledges that the interactions between Lemoine and LaMDA are rather striking.[9]

Some lateral thinking may be helpful here. Somehow related to this is the fact that when talking about AI many of us succumb to the temptation to lump together a range of very different things (Hunter 2023). So far, I have done so myself here. But even if for a moment we limit our consideration to generative AI systems, I have the impression we are facing a variety of different activities and behaviours, which we do tend to think of as bound together, although their mutual link is tenuous. Perhaps this inclination is because the range of “intelligent” activities of generative AI systems consists or includes or is based on linguistic productions in human natural languages (or the production of human-originated images, etc.)? This fact makes the constellation of goal-oriented etc. behaviours of a generative AI system appear close enough to those familiar bundles that we routinely regard as units – human minds, as we call them – for it too to be recognised as a unit, one centre of intelligence.[10] At the same time, its terrific computational power and the impossibility for us to make sense of its procedures and choices makes it also alien enough to be regarded not just as one, but indeed as “the Other”. Or in other words: that AI manifestations are put together, even though in an unusual and perplexing way, from human products, activities and behaviours, means that they are strange projections of our own intelligence; therefore, we tend to recognise them as the quintessential model of an alien

[7] In this connection it would, I think, be worth thinking of other ways in which the issue of AI dovetails with that of embodiment – for example in transhumanism and in various responses to the concept (and realities) of cyborgs (see *e.g.* O’Connell 2017).

[8] Cf. Darwiche 2018, 7: “I believe that attributing human level intelligence to the tasks currently conquered by many neural networks is questionable, as these tasks barely rise to the level of abilities possessed by many animals”. (Darwiche notes, among other things, how recently the threshold for us to be impressed with AI performance has gone down substantially, 5ff.)

[9] As indeed they are. See *e.g.* the interview in Levy 2022, and even more so https://www.youtube.com/watch?v=NAihcvDGaP8; see also https://www.youtube.com/watch?v=XkSu1cWokYA&t=7s and https://www.youtube.com/watch?v=5jaSiROmRV4 (all accessed 14 April 2023).

[10] Worth noting that this is directly contrary to what years ago McCarthy was suggesting would make sense (see above).

intelligence of the kind we were naïvely hoping (*Close Encounters*) or fearing (*Mars Attacks!*) to come across in extra-terrestrial intelligent life.

5. Mindlessly creative?

“We aren’t dealing with ordinary machines here. These are highly complicated pieces of equipment, almost as complicated as living organisms. In some cases, they’ve been designed by other computers. We don’t know exactly how they work”. Thus Michael Crichton in his sci-fi novel *Westworld*, published in 1973.

Connected with the issue of defining “intelligence”, AI experts also seem to make use of specific conceptions of “creativity” and “discovery”. Some AI systems have the ability to instruct themselves beyond the ability of their creators, or to solve problems in ways that their creators or we (humans) do not, or cannot, understand. This is commonly taken to mean that they are being “creative”.[11]

And here is where the troubles start. Many of us wonder, are they really? The question is, as ever, what would we all be willing to count as evidence for AI’s creativity, for its discovering something new? As always with AI, from the Turing test onward, there’s the impression that we keep shifting the goalposts. Or are “creativity” and “discovery” by definition what a machine, including AI systems, cannot do?

But does any AI feat constitute an equivalent to (re)producing an actual discovery? Off the top of my head, perhaps if we do not understand how an AI system is proceeding then we can hardly assess whether their procedures are creative or not? Or perhaps have we given up altogether on understanding and explaining, in favour of an extreme form of pragmatism? More trivially, do we take it that there is something “creative” going on precisely *because* we do not understand it?

Here is how Daniel Dennett put it, a (relatively) long time ago, in relation to “intuition”:

> Remember the unresolved question of how Gary Kasparov uses his “intuitive” powers to play chess. Whenever we say we solved some problem “by intuition”, all that really means is *we don’t know how* we solved it. The simplest way of modeling “intuition” in a computer is simply denying the computer program any access to its own inner workings. Whenever it solves a problem, and you ask it how it solved the problem, it should respond: “I don’t know; it just came to me by intuition”. (Dennett 1997, 29)

In this at least, current AI systems are just like us: black boxes. Is that the mark of intuition or creativity?

In sum, first we underwrite without further question the idea that creativity, intuition and the imagination, and with them discovery, are quintessentially and by definition irrational and incomprehensible; and then we conclude, fallaciously as it happens, that some procedure is

[11] That our failure to understand the operation of AI systems is a sign of their creativity is taken for granted practically everywhere. For examples that are very far apart in scope and tone, see the attacks on Chomsky *et al.* 2023, such as Aaronson 2023 and responses; and Cristianini 2023, 7, 57, etc.

"creative" simply on the grounds that we do not comprehend it? Shouldn't we think a bit harder, indeed do some honest research, in the reasonable hope to reach a better understanding of AI systems' ways of doing business?

It is obvious that I must be missing something here; I'd better try harder. Here is an example of what is meant: while accessing an old programme and playing the resulting computer game, an AI agent found a way to pile up lots of points by making moves that appear to expert human observers not to follow the rules, in fact to be random. This is the story of the old computer game Q*bert in the so-called "Freiburg experiment". As Cristianini explains, the reason for this is that the AI agent uncovered an unintended consequence, in the old game's programme, of tiny shortcutting choices on the part of the original programmer which a human player would be wildly unlikely ever to come across. Hence the surprise effect (Cristianini 2023, ch. 8). And here is an uncannily apt remark by Alan Turing:

> Machines take me by surprise with great frequency. This is largely because I do not do sufficient calculation to decide what to expect them to do, or rather because, although I do a calculation, I do it in a hurried, slipshod fashion, taking risks. (Turing 1950, 450)

Similarly, but with added complexities and important consequences heightening one's attention, consider the recent discovery of the antibiotic Halicin by a group of MIT researchers with the assistance of a neural network. After training it on a couple of thousand molecules known to inhibit the growth of *Escherechia coli*, having the results of the training checked on a larger database, etc., the researchers unleashed the neural network on a database of 107 million molecules, to single out those likely to exercise an equally inhibiting effect. The surprise was not that the neural network did identify some promising molecules, and in particular the one they named Halicin; but that the mechanism through which Halicin operates was, until that moment, unknown – *i.e.*, presumably, unobserved, at least by humans, in any of the molecules used to train the neural network in question (see Stokes *et al.* 2020; Marchant 2020).[12] Do these count as "discoveries" or not?

One would certainly be tempted to answer in the negative in the case of the Freiburg experiment, for the lack of a supporting theory makes that look at best like a case of trivial blasting one's way through randomly, rather than an act of creativity (at least in any ordinary sense – or are we redefining the term?). In fact, Q*bert's behaviour, with its Golem-like inflexible focus on just the one goal of increasing point-collection, looks like a mindless (oops) caricature of creativity and discovery, or indeed of intelligence as it is commonly intended when we are not talking about AI. And yet, the AI agent's move did undeniably bring to light something that was not in evidence before – the programming shortcuts and their unpredicted effects. And it also does seem to count as unambiguously intelligent according to Albus' definition of "intelligence".

[12] This methodology is, unsurprisingly, very fruitful: see *e.g.* Liu, Catacutan, Rathod *et al.* 2023 for a more recent and equally impressive discovery, by the same group, of Abaucin, another antibiotic that promises to be very effective in narrowly targeting a superbug.

The case of Halicin, for all its added complexity and its impressive breakthrough in an area that seriously profited from it, seems to me very similar in its basic structure. And yet, in commenting on it Kissinger *et al.* note that the humans setting the AI system's task appear to have been unable, even *a posteriori*, to work out how the AI system identified that particular molecule as the right one to fulfil the task:

> the AI did not just process data more quickly than humanly possible; it also detected aspects of reality humans have not detected, or *perhaps cannot detect* (italics mine; Kissinger *et al.* 2022, 11).

The question is: the current predicament created for humans by this kind of AI's *modus operandi* is a crisis of our world, confined as it is within the Kantian "bounds of sense", as intimated by Kissinger? *I.e.*, do we have here an intelligence for whom empirical reality is, apparently, dramatically different from ours and simply inaccessible to us in principle? This would not be entirely surprising, if its "embodiment" is also so dramatically different from ours – for example, would this kind of embodiment have any reason to attribute a special place to three-dimensional space…? Or is it the case that this impossibility is due not to any qualitative difference, but just to the massive disparity in processing power? If the latter, perhaps does a lot of quantity end up making a qualitative difference? And even so, should we not start to work hard to map what new bounds – clearly different from those that apply to our senses – do or could apply to AI systems?

In sum, perhaps theories are, after all, no more than guesswork abridgments for use on the part of minds with a sadly limited processing capacity. "Intelligence for a system with limited processing resources consists in making wise choices of what do to next" (Newell and Simon 1976, 98): this certainly obtains for human intelligence but at least apparently, at least in practice, and at least for the time being – *pace* Newell and Simon – not for AI. But this state of affairs, I submit, may or not last very long.

6. Directions of travel?

According to some, reasoning about causation may give the discussion a different turn: "To build truly intelligent machines, teach them cause and effect" (Pearl 2019). If so, the next question is: to what extent can this be done? There have been recent developments in this area, although the combination of much larger availability of data and much higher computational capacity, with its flashy results, seems to have overshadowed the need for this (see Darwiche 2018).

Pearl 2019 is a reader-friendly popularisation; but this line of investigation gets rather technical rather fast, and this is the kind of subject in which the devil is indeed in the details. Nonetheless I shall venture a guess: given the widely voiced concerns about the opacity of process in purely data-driven AI systems, their bluntness and openness to easy bias, etc., not to mention the cost and the environmental concerns that are starting to be raised over the phenomenal energy consumption needed for them, I would not be surprised if there was soon a serious surge of interest in research for increased efficiency – *i.e.*, clever ways to reduce the deployment of the available processing capacity, such as, for example, "causal AI".[13] Unsurprisingly, there is already widespread attention to this in business consultancies.

Hence my next question, that is a very Humean and classic philosophical problem: does a statistical approach to causation manage to account for it without residue? To put it otherwise: is causal understanding to be fully reduced to association (seeing regularities), intervention (predicting and doing) and counterfactuals (imagining, theorising), *i.e.* to a very sophisticated statistical treatment of correlations and counterfactuals (Pearl 2019, 27)?

But in fact, perhaps our wish for more metaphysically robust causal explanations is itself a problem? And after all, for all the prejudices I have due to my philosophy of science background, when talking about AI it may not be prudent to assume any symmetry between explaining and predicting, between understanding how a system works and being able to anticipate how it will behave. So perhaps rather than looking for explanations, we should focus our attention (and work hard) on the behaviour of AI systems and on its predictability (see *e.g.* Amigoni and Schiaffonati 2021). As in the eighteenth century someone wise put it about differential calculus, may this be another case of *allez en avant et la foi vous viendra*?

7. *Stat rosa pristina nomine...?*

An observation in passing on another important language-related aspect of all this. It appears that humans, when faced with AI interpretations of data which suggest a certain course of action, tend to go along with that course of action even against their own better judgment, even though, in fact, often they would be right and the AI wrong (see Hunter 2023, esp. in medical contexts; Hunter presents this as another of the most common temptations we should learn how to avoid when dealing with AI). What is it that makes AI responses and suggested courses of action so compelling? I suggest that this may be accounted for by the combination of two factors. The first is that, as mentioned above, this kind of AI systems uses human language and, in this sense, inevitably gives the impression of being very akin to us. The second is that, at the same time, we tend to think of them as dispassionate and therefore capable of a more objective view – a "view from nowhere".

This latter interests me. I wonder if it is a legacy, perhaps, of our habit to be more easily persuaded by the (apparent) lack of rhetoric of the "writing degree zero" (Barthes 1990) and the objectivity that is typical of the language of science – linked as it is to our fascination with quantitative assessments and with numbers?

It is worth noting that there are odd resonances between the linguistic activities of generative AI systems and aspects of post-modernist literary criticism. Consider Searle's take on "electronic brains" and "artificial intelligence" in terms of intentionality (lack thereof):

> [...] the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbo-

[13] As well as Pearl 2019, see the already cited Darwiche 2018; Sgaier *et al.* 2020 for discussion of some practical examples (climate change, two healthcare issues).

> lize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output […] the programmed computer does not do "information processing". Rather, what it does is manipulate formal symbols. The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. (Searle 1980, 422, 423)

Whether Searle's argument works or begs the question (I think obviously the latter, see *e.g.* Churchland and Churchland 1990 or McCarthy 2001) is not my point here; what I am interested in is the infallible diagnostic clarity with which he presented what he thought was the AI predicament – the view from nowhere as the view without a viewer – and its similarity to the condition so brilliantly sketched in Roland Barthes's metaphor of the death of the author:

> it is language which speaks, not the author; to write is, through a prerequisite impersonality […] to reach that point where only language acts, "performs", and not "me". (Barthes 1977, 143)

This has been read as suggesting that all there is of the world itself is language and text – all discourse is a citation or a tissue of citations from a former discourse that's always already existed. *Stat rosa pristina nomine*, all of it.

In fact, I think this view of generative AI systems' literary productions is a revealing travesty of a Barthesian (however dead) author. Searle thinks that with AI we have speakers or authors who have always already been unable to appropriate language, even though they do use it; and that is because, having always already failed to exist as minds, they do not even need to undergo their own death.

But think of Mikhail Bakhtin's less aphoristic presentation of language ownership:

> The word in language is half someone else's. It becomes one's "own" only when the speaker populates it with his own intentions, his own accent, when he appropriates the word, adapting it to his own semantic and expressive intention. Prior to this moment of appropriation, the word does not exist in a neutral and impersonal language (it is not, after all, out of the dictionary that a speaker gets his words!), but rather it exists in other people's mouths, in other people's contexts, serving other people's intentions: it is from there that one must take the word, and make it one's own. […] Language is not a neutral medium that passes freely and easily into the private property of the speaker's intentions; it is populated – overpopulated – with the intentions of others. (Bakhtin, Holquist and Emerson 1981, 293-94)

I should perhaps confess that my own personal experience as an adult learner of languages is, at least to an extent, in line with this view: I still remember exactly from where I have taken some turns of phrase or idioms ("to get rid of", for example, or "to get to grips with"). I know this is not just my peculiarity. And it is true also that these stratifications in my memory give to some of what I say a half-forgotten feel that, although elusive, does make a difference. So then, *nomina nuda tenemus*…? But in fact, even though the world seems very remote from all this, there is no implication that it does not exist or that all there is of it is words. Nor is there any serious implication that language exists and speaks independently of any speakers or authors. It is, in fact, the exact opposite: language is described as a huge repository expressive of innumerable former and contemporary speakers' contexts, mouths and accents, intentions – that is, their lives. And it is nobody's property, being freely available for any new speaker's appropriation, that is, open to their participation and contribution. This also means, a new speaker's utterances include and are affected by the language's former and contemporary history of utterances and coloured by them. In all these respects Bakhtin's language is, in fact, not very different from the repositories of data on which AI systems are trained.

Incidentally, while Bakhtin looks at this in the frame of reference of language communities, we can also think of it as relating to the visual, and in the case of an individual organism. Thus Nelson Goodman:

> The eye comes always ancient to its work, obsessed by its own past and by old and new insinuations of the ear, nose, tongue, fingers, heart, and brain. It functions not as an instrument self-powered and alone, but as a dutiful member of a complex and capricious organism. Not only how but what it sees is regulated by need and prejudice. It selects, rejects, organizes, discriminates, associates, classifies, analyses, constructs. It does not so much mirror as take and make; and what it takes and makes it sees not bare, as items without attributes, but as things, as food, as people, as enemies, as stars, as weapons. Nothing is seen nakedly or naked. (Goodman 1968, 7-8)

This is an organism that, like an anthill (as in Hofstadter 1979) already works a bit like a society.

8. Social machines

In discussions of AI's superhuman power, the emphasis is always on the sheer amount of stuff AI systems can digest and base their responses on. So, what makes AI "more clever" (more effective, and so more intelligent, although I would hesitate to call them more efficient) than humans is their ability to process disproportionate amounts of info that no human would manage to master in umpteen hundreds of years, etc. This is obviously true of individuals, and indeed this is at the basis of how AI systems have been able to defeat chess champions (and each other, etc.).

But it may perhaps be the case that in this context the relevant comparison is not with the contents, actual or possible, of an individual human mind, but with the overall cumulative mental contents of at least a group of communicating individuals with a range of different areas of expertise, etc.

For quite some time now scientific articles have been signed by large numbers of authors; and in any case the overall knowledge patrimony of a society (or indeed of mankind) is measured by the sum of the various forms of expertise represented in it via subsets of its population; etc. Moreover, the entire business of the division of epistemic labour, knowledge by testimony and reliance on experts may be read as indicating something of this kind: perhaps it is not me, but my species or society that know (say) game theory or cybernetics, and through experts in a sense so do I, although directly and personally I do not?

An anthill is a complex system that displays apparently rational, goal-oriented emerging behaviour. In this sense it is intelligent according to the definitions above. Same for, say, market economy (think of Adam Smith's invisible hand). The market is a "social machine", and one where some (not all) components are humans; and, like the anthill, it is an "intelligent agent", since it does appear to give rise to free and spontaneous purposeful behaviour. Note that, again just like the anthill, it typically does so at the macrolevel and in ways that are above and beyond or out of the control of its components, with the purposes its (human) components selfishly pursue at the microlevel nonetheless cumulatively delivering a macrolevel purpose that is not necessarily in line with them – indeed, the macro- and microlevels may even be and actually often are misaligned, without necessarily causing problems (Cristianini *et al.* 2023, 93, who inevitably cites Hoftstadter 1979 on this). All this also applies to *Wikipedia*, "citizen science" programmes and the Internet of Things,[14] and indeed to all crowd-sourced apps, social media, and various platforms. In fact, the old concept of "social machine" seems to have acquired theoretical momentum in connection with the origin of the internet, and, typically, "social machines" as now newly defined (here we go again) involve a mix of humans and AI. And now, what is needed is ways to mitigate the threat they may pose to individuals' autonomy in a very basic sense (*ibid.*, 94: think of, *e.g.*, micro-nudging or collection and exploitation of psychometric info, not to mention obvious employment issues, etc.).

9. An alliance against trolls

> […] some cognitive tasks can be emulated to a reasonable extent without the need to understand or formalize these cognitive tasks as originally believed and sought (as in some speech and vision applications). That is, we succeeded in these applications by having circumvented certain technical challenges instead of having solved them directly. This observation is not meant to discount current success, but to highlight its nature and lay the grounds for the following question: How far can we go with this direction? (Darwiche 2018, 6)

A mere handful of years down the line, we can already tell that apparently we can go rather far; this is, in any case, the widespread public perception now.

The fact that the discussion of all this is now very public is a consequence of just how far we have got and how fast; and that it tends rapidly to become so acrimonious is, I think, unsurprising – but interesting. In particular, recently Chomsky was targeted as the prize representative of the AI that was dreamt up originally (see *e.g.* Aaronson 2023 and comments). His was an AI based on understanding, explaining and modelling human intelligence, like McCarthy's,

[14] Key features of *Wikipedia*, citizen science programmes, IoT as social machines are listed in N. Shadbolt *et al.* 2016, 111.

like Simon's, all of them in their different ways. He is a surviving expert, now in his 90s, of a research programme that they concluded with an open acknowledgement that it was simply impossible for them to deliver their intended big prize. That negative result was a key discovery in its own right. Now the new experts seem to forget their debt to them and to glory in the brute force of their new toys' *modus operandi*, in their own exhibited anti-intellectualism, and alas, also in the opportunity to hurl abuse at the defeated hero of a different, more intellectually fastidious way to go about it. At times brave new AI, for all that it achieves an appearance of cleverness in its effectiveness, allows for attitudes and behaviours that are not pretty, just like the mass society expressing it.

To get back where I started: the unusual meanings of "intelligence", "creativity", "discovery", etc. are all derived from a range of specialised aspects of computer sciences, engineering, economics, sociology (AI, control theory, game theory, theory of choice, behavioural economics and nudge theory, STS and ANT, etc.). Among the practitioners of these disciplines and the AI experts piggybacking on their terminology all this is well-known and taken for granted. But now that AI-related enthusiasm and concerns alike have become a major topic of conversation in newspapers, magazines and social media, non-specialists are taking a keen interest and joining into discussions; and for them these technical concepts and terminology are, for the most part, a novelty. So it is no surprise if there is confusion: participants in the debate literally talk at cross-purposes to each other, and tempers get hot, as they always do very rapidly in such debates.

There is rather a bleak world that one can easily conjure up in this way, and some are, inevitably, doing so. This is where, I think, closer cooperation between AI and philosophy researchers may well help to bring about a much-needed new humanism.

References

Aaronson, Scott. 2023. "The False Promise of Chomskyism". *Shtetl Optimized. The Blog of Scott Aaronson*, March 10. https://scottaaronson.blog/?p=7094.

Albus, James S. 1991. "Outline for a Theory of Intelligence". *IEEE Trans. Syst., Man Cybern.* 21, no. 3 (May/June): 473-509.

Amigoni, Francesco, and Schiaffonati, Viola. 2021. "The Importance of Prediction in Designing Artificial Intelligence Systems". In *Machines We Trust*, edited by Marcello Pelillo and Teresa Scantamburlo, 105-119. Cambridge, MA: MIT Press.

Bakhtin, Michael, Holquist, Michael, and Emerson, Caryl. 1981. *The Dialogic Imagination: Four Essays by M. M. Bakhtin*. Austin: University of Texas Press.

Barthes, Roland. 1977. "The Death of the Author". In *Image, Music, Text*, edited by S. Heath, 142-148. London: Fontana.

Barthes, Roland. 1967. *Writing Degree Zero*. London: Jonathan Cape.

Burr, Christopher, Cristianini, Nello, and Ladyman, James. 2018. "An Analysis of the Interaction between Intelligent Software Agents and Human Users". *Minds Mach.* 28: 735-774.

Chalmers, David. J. 2022. "Can a Large Language Model Be Conscious?". https://arxiv.org/pdf/2303.07103.pdf.

Chomsky, Noam, Roberts, Ian, and Watumull, Jeffrey. 2023. “The False Promise of ChatGPT”. *The New York Times*. March 8, 2023. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Churchland, Paul M. and Churchland, Patricia S. 1990. “Could a Machine Think?”. *Sci. Am.* 262 (1): 32-39.

Contucci, Pierluigi. 2023. *Rivoluzione Intelligenza Artificiale. Sfide, rischi e opportunità*. Bari: Edizioni Dedalo.

Cristianini, Nello. 2021. “Shortcuts to Artificial Intelligence”. In *Machines We Trust*, edited by Marcello Pelillo and Teresa Scantamburlo, 11-25. Cambridge, MA: MIT Press.

Cristianini, Nello. 2023. *The Shortcut. Why Intelligent Machines Do Not Think like Us*. Abington, Oxon: CRC Press.

Cristianini, Nello, Scantamburlo, Teresa, and Ladyman, James. 2023. “The Social Turn of Artificial Intelligence”. *AI & Society* 38: 89-96.

Darwiche, Adnan. 2018. “Human-Level Intelligence or Animal-Like Abilities?”. *Commun. ACM* 61 (10): 56-67.

Dennett, Daniel. 1997. “Can Machines Think? Deep Blue and Beyond”. *Studium Generale Maastricht*: 10-32.

EU AI High-Level Expert Group (AI HLEG). 2019. *Ethics Guidelines for Trustworthy AI*, 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

Goodman, Nelson. 1968. *Languages of Art: An Approach to a Theory of Symbols*. Indianapolis: The Bobbs-Merrill Company.

Hofstadter, Douglas. 1979. *Gödel Escher Bach. An Eternal Golden Braid*. Harmondsworth, Middlesex: Penguin Books.

Hunter, Tatum. 2023. “Three Things Everyone’s Getting Wrong about AI”. *Washington Post* March 22, 2023. https://www.washingtonpost.com/technology/2023/03/22/ai-red-flags-misinformation/.

Jacquette, Dale. 1989. “Adventures in the Chinese Room”. *Philos. Phenom. Res.* 49: 605-623.

Kissinger, Henry, Schmidt, Eric, and Huttenlocher, Daniel. 2022. *The Age of AI*. London: John Murray.

Levy, Steven. 2022. “Blake Lemoine Says Google’s LaMDA AI Faces ‘Bigotry’”. *Wired* June 17, 2022. https://www.wired.com/story/blake-lemoine-google-lamda-ai-bigotry/.

Liu, Gary, Catacutan, Denise B., Rathod, Khushi, *et al*. 2023. “Deep Learning-Guided Discovery of an Antibiotic Targeting *Acinetobacter baumannii*”. *Nat. Chem. Biol.* 25. https://doi.org/10.1038/s41589-023-01349-8.

Marchant, Jo. 2020. “Powerful Antibiotics Discovered Using AI”. *Nature* February 20, 2020. https://doi.org/10.1038/d41586-020-00018-3.

McCarthy, John. 1979. “Ascribing Mental Qualities to Machines”. http://jmc.stanford.edu/articles/ascribing/ascribing.pdf.

McCarthy, John. 2001. “John Searle’s Chinese Room Argument”. http://jmc.stanford.edu/articles/chinese.html.

McCarthy, John. 2002. “Simple Deterministic Free Will”. http://www-formal.stanford.edu/jmc/freewill2.pdf.

McCarthy, John. 2006. "What Has AI in common with Philosophy?". http://www-formal.stanford.edu/jmc/.

Newell, Allen, and Simon, Herbert A. 1976. "Computer Science as Empirical Inquiry: Symbols and Search". *Commun. ACM* 19 (3): 113-126.

O'Connell, Mark. 2017. *To Be a Machine: Adventures Among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death*. London: Granta Books.

Pearl, Judea, Mackenzie, Dana. 2019. *The Book of Why. The New Science of Cause and Effect*. London: Penguin Books.

Searle, John R. 1980. "Minds, Brains, and Programmes". *Behav. Brain Sci.* 3: 417-457.

Sgaier, Sema K., Huang, Vincent, and Charles, Grace. 2020. "The Case for Causal AI". *Stanf. Soc. Innov. Rev.* 18 (3): 50-55.

Shadbolt, Nigel, Van Kleek, Max, and Binns, Reuben. 2016. "The Rise of Social Machines". *IEEE Consum. Electron. Mag.* 5 (2): 106-111.

Stokes, Jonathan M., Yang, Kevin, Swanson, Kyle, *et al*. 2020. "A Deep-Learning Approach to Antibiotic Discovery". *Cell.* 180: 688-702.

Turing, Alan M. 1950. "Can Machines Think?". *Mind* 59 (236): 433-460.

Veruggio, Gianmarco. 2006. "The EURON Roboethics Roadmap". In *Proc. Humanoids '06: 6th IEEE-RAS Int. Conf. Humanoid Robots*, 612-617. doi: 10.1109/ICHR.2006.321337.

https://www3.nd.edu/~rbarger/ethics-roadmap.pdf.