The new ghosts in the machine: 'Pragmatist' AI and the conceptual perils of anthropomorphic description

Phillip Brooker
University of Liverpool

William Dutton
Peak, Manchester

Michael Mair
University of Liverpool

Abstract

Algorithms are becoming interwoven with increasingly many aspects of our affairs. That process of interweaving has brought with it a language laden with anthropomorphic descriptions of the technologies involved, which variously hint at 'human-esque' or 'conscious-like' activity occurring within or behind their operations. Indeed, the term 'Artificial Intelligence' (AI) seems to refer to a quality that is thought to be largely human; namely, intelligence. However, while anthropomorphic descriptions may be useful or harmless, when taken at face value they generate a false picture of algorithms as well as of our own thinking and reasoning practices by treating them as analogues of one another rather than as distinct. Focusing on the algorithm, and what it is misleadingly said to be and to be like, in this article we outline three 'perspicuous representations' (Wittgenstein 1953: §122) of AI in specific contexts. Drawing on Wes Sharrock's ethnomethodological and Wittgensteinian work, our aim is to demonstrate that by attending to the particular, occasioned and locally accountable, not to say highly specified, usages of language that accompany the 'New AI' in particular, we can avoid being haunted by the new task performing ghosts currently being discursively conjured up in our algorithmic machines.

INTRODUCTION

> If calculating looks to us like the action of a machine, it is the *human being* doing the calculation that is the machine. (Wittgenstein 1956: IV; §20, 234)

Algorithms are becoming interwoven with increasingly many aspects of our affairs. That process of interweaving has brought with it a language laden with anthropomorphic descriptions of the technologies involved, which variously hint at 'human-esque' or 'conscious-like' activity occurring within or behind their operations. Indeed, Artificial Intelligence (AI), as the designator that groups together work in this contemporary interdisciplinary field of science, technology and engineering, seems to refer to a quality that is thought to be largely human; namely, intelligence. Within that broad field of AI there are also many other popular designators within its various subdomains that evoke anthropological qualities too; such as those we find in references to 'machine learning', 'deep learning', 'neural nets', 'decision trees', in Google's naming of its 'DeepDream' algorithm, and so on. All share the same surface grammar, suggesting activity taking place in the workings of algorithms in some way akin to human thought and reasoning.

While these anthropomorphic descriptions may be useful or indeed harmless, when taken at face value (as in academic treatises that conflate computational processes with human minds through a cognitivist/materialist/behaviourist lens most prominently those of major figures such as Fodor (1981), Stich (1983) and Dennett (1987) working in the tradition of analytic philosophy and the Theory of Mind)[1] they generate a false picture of algorithms as well as our own thinking and reasoning practices by treating them as analogues of one another rather than as distinct. That strained analogical equivalence is often established by metaphors that seem to provide a bridge between both: for example, 'the mind is a machine; the computer is a machine; therefore, the mind and the computer are alike' (cf. Brockless 2019). While the claim that the mind is a machine could itself be written about at great length, in line with Wes Sharrock's work we are more interested here in the other side of the equation; namely the algorithm, and what it is misleadingly said to be and to be like.[2] This is because, in addition to distorting

[1] This is an incomplete list, of course, and could extend to a very much wider range of figures working on AI in the analytic tradition from Turing onwards. Given our interest in what has been termed New AI, we will not extend it further here except to note the locus of the problems has shifted but the misconceptions are of a related kind. Instead, we would direct readers to the work of Sharrock and Coulter (e.g., 2004; 2009) and Coulter and Sharrock (2007) alongside the other works by Sharrock and colleagues we cite below for a much fuller account of analytic philosophy's conceptually misconceived Theory of Mind.

[2] While this is a subject of growing interest in contemporary social science as elsewhere (e.g., Mackenzie 2015; Burrell 2016; Stilgoe, 2018), even the researchers most sensitive to the problems find it hard to free themselves fully from the baleful conceptual legacy passed on by analytic philosophy's formal account of both mind and artificial intelligence as one of its expressions, something

understandings of thinking and reasoning, the computational or information processing model of mind also produces distorted understandings of how algorithms work and what gives them life—namely the practical uses to which people intendedly put them in particular contexts. Alive to these issues, contemporary figures in the 'New AI' have sought to eschew the vocabulary of cognitivism and thus avoid its associated (metaphysical) problems, disavowing any suggestion they are modelling algorithmic 'minds' on an information-processing or, indeed, any other kind of theoretical model, preferring to talk instead of their creations as 'problem-solving' technologies.[3] As Runciman notes, 'one of the reasons for the remarkably rapid recent progress in machine learning [as a currently-prominent subset of AI] has been the deliberate detachment of algorithmic problem-solving from hoary questions about what counts as true intelligence' (2018: 38). Nonetheless, despite these disavowals of AI's cognitivist residue, our recurrent tendency to ascribe agency and intentionality to the inanimate, examined by Wittgenstein in his (1959) *Remarks on Frazer's Golden Bough* for instance, appears to be particularly and stubbornly persistent when it comes to describing AI's contemporary artefacts.

The tendency to loosely grant AI the status of some kind of synthetic consciousness seems (and often is) relatively banal, but it can lead to significant confusions—the conjuring up of new 'task performing' or 'agentic' ghosts to be installed in our algorithmic machines (Ryle [1949] 2000). These ghostly spectres, projected by misconceptions of AI, can have practical consequences. They can, for instance, lead to the allocation of resources to ill-designed research and poorly-thought-through applications, or potentially more pernicious problems like the shirking of moral responsibility in the use of AI. For example, as we will go on to discuss, such an accusation can be levied at Microsoft's 'Tay' chatbot, which 'learned' to tweet (or rather was not designed to filter out instructions to reproduce) racist content when left in the wilds of Twitter, unsupervised, by its creators (cf. Perez

particularly visible in discussions of the 'learning' done in 'machine learning' as we shall go on to discuss—see also footnote 3 below.

[3] Demis Hassabis is probably the most prominent proponent of the New AI today thanks to his leadership of Google DeepMind, the Google offshoot responsible for the AlphaGo algorithm, which famously 'beat' the human world champion, Lee Seedol, at Go in a five-match series in 2016. The titles of two of his team's papers on AlphaGo indicate the tension we are seeking to highlight and analyse here; the papers, 'Mastering the Game of Go with Deep Neural Networks and Tree Search' (Silver et al. 2016) and 'Mastering the Game of Go without Human Knowledge' (Silver et al. 2017). With close to 10,000 citations between them, we can see a shift from the earlier to later paper where the technical talk of 'deep neural networks and tree search' reverts to the simpler 'without human knowledge'. This reflects Hassabis and colleagues' position that AlphaGo does not know how to play Go or even that it's playing, a hallmark of the New AI. However, they persist in talking of 'mastery'. This is the nub of the matter for us. Hassabis and colleagues undoubted engineering achievements are not something we wish to bring into question; instead, we argue that they slip out of engineering and into philosophy when they use anthropomorphic descriptions to suggest the algorithm's performance can be treated as 'mastery' of Go.

2016). As a prophylactic against those confusions and their repercussions, our intention here is to outline a series of three 'perspicuous representations' (Wittgenstein 1953: §122) of AI in specific contexts, drawing on both Wes Sharrock's ethnomethodological and Wittgensteinian work (e.g., Button, Coulter, Lee and Sharrock 1995; Sharrock and Button 1999; Sharrock and Coulter 2004) in order to demonstrate how we can avoid being haunted by these new ghosts and 'their' shadowy work.

By way of these perspicuous representations—the aforementioned case of Microsoft's Tay chatbot, plus reflection on text conversations between staff in a commercial consultancy for AI applications, as well as observations on two crude pieces of (Python) code designed to test the limits of how we might talk about AI in its native guise and habitat—we aim to show that an algorithm's production is dependent on a great deal of 'scene setting' that has little if anything to do with simulated 'intelligence' or computational 'agency'. As we will demonstrate, the practices and accounts which set the scene for 'intelligent technologies' cover a broad sphere of activity in an algorithm's assembly, from coding, computer science, mathematical and statistical work, to understandings of the context in which they are to be put to use, all of which are far removed from activity that could be accurately described as 'creating a synthetic consciousness'. What our examples will show is that the performance (or 'learning') of an algorithm is not determined (solely) by its performance mathematically (though that is important), but by the 'occasioned character' (Sharrock and Ikeya 2000: 275) of the activities it is bound up with (including language-use) and the practical differences it makes to the situations in which it is or will be put to use. Furthermore, our examples will do so by examining the ways in which that difference is designedly made accountable—in Sharrock and Anderson's terms, 'how a body of knowledge and the courses [of action and reasoning] associated with it, can be viewed as organised to be found, to be used, to be understood, and how the use and understanding of such knowledge[, action] and reasoning is the display of its organisation' (2011: 47). As such, the ascription of 'psychological' or 'cognitive' predicates to AI technologies will be demonstrated to be a contextually embedded, occasioned and thus locally accountable matter, with algorithms only capable of being described as 'intelligent' in circumscribed ways within the structures of practical activity in which they play a part.

Crucially, such particular, occasioned and locally accountable, not to say highly *specified*, usages of the term 'intelligence' are *not* the same as those we find in philosophical or conceptual works on AI—and our perspicuous representations are intended to draw attention to precisely this difference between metaphysical formalism and the work-a-day. Pulling the strands of our argument and analysis together, in conclusion we suggest the best way to resist conceptual confusion in our treatments of AI is to remain resolutely fixed on the work done (or perhaps, *not* being done) with anthropomorphic descriptions and thus the wider sets of

practices—the localised 'forms of life' and their 'surroundings' (Wittgenstein 1957: Part VII; §47, 413)—this technology is interwoven with and finds (or fails to find) practical uses within. Attending to such matters, we argue, lays the conceptually conjured ghosts to rest by showing there is no need to ascribe new spectral agencies to our contemporary computational tools. Stated baldly, our ways of talking about and doing things with AI technologies commit us to no particular metaphysical positions at all.

ANTHROPOMORPHICALLY DESCRIBING AI

AI is an increasingly familiar cultural object. Though AI *does* still form the basis of a great deal of speculative science fiction, AI and related technologies are spread ubiquitously throughout social life; 'it' can be found in our workplaces, 'it' features in our political deliberations, 'it' is a deep part of our social media networking, 'it' shows up in our internet shopping in the things we buy and how we buy them, and much more besides. In short, AI is no longer (just) a recurrent motif in mostly apocalyptic visions of our future, but a mundane and ordinary part of our present-day lives. AI has been developed and has found application across all manner of contexts and domains, with the effect that we are seeing an increasingly varied array of terms for 'it' being brought into play:[4] 'artificial intelligence' in the broad sense is now also to be found as 'machine learning' mechanisms in workplaces involved in data analytics, as 'deep learning' in academic research, 'neural nets' in image classification software on social media networks, 'decision trees' in business management, and so on. It is possible to talk about AI as more than these three orders of technological innovation—as more than a field (AI), an approach (machine learning) or set of mechanisms (specific algorithms or applications)—but these orders are not bad places to begin. This is partly because the terms in which they are cast share the surface grammar we briefly sketched above, simultaneously highlighting the engineered nature (and processing power) of the software, alongside a range of activities more normally associated with humans and human consciousness—the 'learning' here being done by 'machines', the 'neural' capacity of the algorithms being 'lodged' in extended and connected 'net(work)s' rather than the 'fleshy' and only 'softly networked containers' that proponents of contemporary AI routinely suggest our 'organic' consciousnesses are 'housed' within (cf. Ryle [1949] 2000 on this container conception of bodies). In short, the 'intelligence' displayed by these technologies is not naturally occurring but 'artificial', and it is regarded as enhanced by being so.

[4] Our focus on language and terminology is indebted to the work of Wittgenstein (e.g., 1953) and Winch (e.g., 1958) and the influence each have had on ethnomethodology, where a focus on language-use in its situated context is held as a rejoinder to the misleading usages of language we find in the local settings of analytic philosophy and much academic social science besides.

As we have suggested, identifying applications of these technologies (and usages of vernacular terms around them) can be seen as a largely trivial task, due to their cultural ubiquity. In various branches of academic and philosophical study broadly taking up questions in and around the Theory of Mind, it is treated as a similarly straightforward task. Yet our use of scare quotes in references to AI as an 'it', is meant to suggest we might be wise not to proceed so casually. While the results of machine learning mechanisms are on display in the returns of any browser search or trawl through a social media feed, and are thus available to 'anyone' (Sacks 1992, vol. 1: 40; Anderson and Sharrock 1979), those machine learning mechanisms, like AI technologies more generally, only occasionally are. A shorthand reference to the 'it' which 'produced' those results makes life easier; it has a certain economy, avoiding the need for lengthy technical descriptions, and thus allows members to 'bracket' the terms of their engagements with AI (Schutz 1962). Few of us would feel bound to discuss internal combustion engines or the physical principles which underpin their operations when discussing our cars, and it would be unwieldy and hence unreasonable to expect the situation to be any different with respect to the AI technologies we now ubiquitously make use of. Indeed, the digital realm is set up so we do not have to explore the machinery unless we really want to—and that is a boon. We can just search or trawl, like we can just turn the keys in the ignition, and thus get going without further ado. Instead of scare quotes, this mundane and ubiquitous bracketing process can be captured through the use of asterisks, something pioneered by Harold Garfinkel (2002), where we can refer to our engagement with some AI device, mechanism or application as an engagement with some particular it*, where the * denotes an as yet to be—and perhaps never to be—finally specified or settled understanding of the asterisked term. Our encounters with the its* of AI, as bracketed engagements, are, as Alfred Schutz (1962) once put it, transparent *enough* for almost all our practical purposes (see also Garfinkel 1967). It is only where things *stop* working transparently that we ever find the need to engage in any deeper considerations, and even then those considerations will tend to be of a relatively 'shallow' sort—'have you logged on using your normal account?', i.e., only as extended as they need to be to enable us to get back to getting on with things.

Problems start to arise when analytic philosophers and those who have followed seek to leave such bracketed engagements behind. Our view, following Button, Coulter, Lee and Sharrock (1995), is that formal philosophical accounts of AI in the analytic tradition produce deep-seated confusions because they take the its* of AI—or, rather, the analytic philosopher's its*—*to be AI*. In other words, while dressed up in seemingly technical language, analytic philosophers trade on understandings of the technicalities of AI which only need to be 'good enough' for *their* practical philosophical purposes. As a consequence, their accounts tell us more about their philosophical preoccupations than the technologies which provide the occasion for elaborating them. Indeed, analytic philosophers are, by and large,

uninterested in the technical specifics of AI—sets of computational operations performed or 'trained' on specified datasets linked step-wise or 'instructed' via lines of code elaborated into executable programmes which build on the results of previous iterations or runs on the 'training' datasets.[5] When setting out to produce their accounts, few philosophers (or social scientists) feel the need to do some coding or programming or even learn enough of the basics to follow what those who are involved in coding or programming might be up to. They don't have to; things are organised so they can get on with their business without needing to. They thus make use of AI in all sorts of unexamined and taken-for-granted ways.[6]

We want, by contrast, to think about a different kind of engagement with AI, one of the sort Wes Sharrock would recommend. That is, we will not argue that we need to get the its* out of the way so we can better see how AI *really* works. For us, that would be, as Garfinkel puts it, 'taking away the walls of the house to see what is keeping the roof up' (as paraphrased in Sharrock 1989: 668). Rather, we feel it is best to come at AI via the its*, and so at the technicalities via the sorts of ordinary, practical engagements where considerations of the technicalities become relevant and accountable in hands-on ways. These sorts of engagements are instructive, providing us with 'tutorials' in AI and its workings connected to specific practical problems and their accompanying troubles (Garfinkel 2002). More specifically, and as part of this, we are interested in situations where members themselves switch back and forth between the displayed results of the technical operations of AI and a consideration of the technical operations themselves. These kinds of situations can tell us a great deal about AI, not least by helping us dispel philosophically generated confusions by furnishing 'reminders' (Wittgenstein 1953) that AI technologies both arise from and are interwoven with particular practices and projects. By refocusing on AI in this way, we hope, among other things, to show that it is perfectly possible to live with anthropomorphic descriptions—'Oh no, why's it doing that now?'—so long as we resist the temptation to read epistemological or ontological, that is, metaphysical, commitments into them (Coulter 1979). All in all, we want to demonstrate that attending to such matters can teach us what we might be talking about when we talk about AI.

[5] While analytic philosophers typically show no such caution, we would be hesitant to call multiplying the results of six rolls of a die together an 'it', and we would extend that hesitancy to descriptions of AI more broadly. A better question, one we go on to consider in more detail below, might be: How are AI technologies put together and where and how does the work of putting them together become relevantly discussable.

[6] Something which would, again, be unproblematic were it not for the claim that generalised accounts of AI capture what AI *really* is.

INTELLIGENCE AT WORK/INTELLIGENCE ON DISPLAY

Discussions of AI technologies often begin with the suggestion that, within them, we find intelligence at work or that they display intelligence *in some sense*. While we feel little compulsion to think of our 'smart' televisions, speakers or phones as anything more than 'smartly' designed, allowing us to talk, wave at or otherwise programme them to do things in various helpfully uninvolved ways, when it comes to AI the claim is often that the technologies we're discussing are more than just 'intelligently' designed. What could perhaps be better taken as a way of delineating *types* of technological innovation has come to be treated as if were freighted with much deeper implications. The claim is that the 'smart' technology is actually quite 'dumb', like most machines, whereas an AI really displays intelligence, albeit again only in *some* sense. In what sense though? In order to answer that question, we want to trace the development of AI as a field back to the Turing Test, and Alan Turing's original delineation of a future vision for computing, following Lucy Suchman's (2019) lead but also that of Stuart Shanker (1987) in doing so.

In reflecting on these issues, we found it useful to remind ourselves of the point that the Turing Test, and even 'Artificial Intelligence' as it was originally conceived in the 1930s, was not seen as having conceptual or philosophical implications per se. Yet that is how those who traffic in generalised conceptions of AI have precisely come to use them—they have taken up these terms and taken them 'curiously seriously' (Garfinkel 1990), assuming that the analogies in play are not merely labels but actually descriptive of the stuff that goes on in the fabric of how AI is built, what it is for, what it does, how it is to be applied, and so on. Thus, the term 'intelligence' is taken seriously through attempts to explain 'intelligence' in humans by equating it with sets of computational processes, making of it an epistemological issue which epistemologists—not engineers, statisticians or the like—are needed to (somehow) resolve. The word 'artificial' has also been taken to have serious philosophical implications because it has been (re-)pitched as an ontological concern about what it means to be 'real' (or human), synthetic or somewhere in between. Once again, when cast in these terms, the problem becomes one for ontologists to (somehow) resolve, i.e., a philosophical not a practical issue.

Neither of these philosophical ways of defining the issues (either alone or in combination) maps onto what was originally intended to be *a*—not *the*—way of describing computing systems where the output is unexpected or unpredictable (by humans) in advance of the input. For instance, according to Chollet and Allaire (2018: 5):

> In classical programming, the paradigm of symbolic AI, humans input rules (a program) and data to be processed according to these rules, and out come answers. With machine learning, humans input data as well as the answers expected from

> the data, and out come the rules. These rules can then be applied to new data to produce original answers.

Accounts of the kind Chollet and Allaire provide might be thought of as more than sufficient to enable us to deflate AI's metaphysical balloon. That has not proven to be the case. Faced with the mundane work of building an AI system and its practical specifics—work in which it is difficult to find philosophical problems—one response has been to argue that the production process is one thing and the product quite another. Can machines not outdo their designers? Are they not capable of going beyond the human? Indeed, might they not be capable of 'teaching themselves' human capacities and thus of developing their own 'agencies'? This is the imaginary bequeathed to us by Turing and his famous Test. But what *was* the Turing Test? It wasn't the demand that someone design a machine that would actually be 'intelligent'. Turing realized that was far too great a task, at least initially. It was instead a way of thinking through whether it might be possible to intelligently design a machine that could fool those who interacted with it into *thinking* it was human by *simulating* intelligence, thus scoring a victory in an 'imitation game' (Turing 1950). Unfortunately, however, Turing's provocation has become something of a fetish,[7] not only among philosophers but also among AI designers, something that can be seen in the contemporary obsession with building programmes that can 'beat' human players at draughts, chess, Go or computer games, despite the fact that the most useful and used (as well as most promising and profitable) applications of machine learning lie elsewhere entirely—as our ubiquitous recourse to search engines among other things shows in its own very ordinary way.

Yet the conceptual slippage that enabled this shift of focus from engineering results good enough to generate what people might mistakenly treat as displays of intelligence to engineering actual intelligence—whatever that might mean—wasn't picked up when it was first made (nobody listened to Wes Sharrock!). This has had the effect of steering the ways in which computational innovations have developed subsequently. It has also influenced how they are now understood and worked with in all sorts of domains. Imagine two people yoked together by a rope, each hoping to travel forward with one another in a relatively harmonious way. If one veers off however, the other is brought with them. Just so, work on AI *and* elementary philosophical mistakes have now become entangled, leading both to lurch off on skewed angles as a result of anthropomorphising AI in the quest to create a bona fide, self-standing, *un*asterisked artificially intelligent 'it'.

---

[7] Indeed, it became so for Turing himself who switched from what might be termed an engineering perspective, i.e., 'How do we make a machine that calculates reliably?', to a philosophical one, i.e., 'Can we build machines that are actually intelligent?', as his work progressed from the 1930s to the 1950s (and on this point see the excellent account by Shanker 1987).

ORDINARY LANGUAGE DESCRIPTIONS OF AI

Conceptualising AI in formal analytic terms not only encourages us to read epistemological and ontological commitments into otherwise harmless anthropomorphic descriptions. The problem is wider. We all (that is, everybody; not just philosophers) have a fairly poor understanding of the roles of computational innovations in our daily lives,[8] and we can readily confuse wildly speculative sci-fi for probable near-future scenarios. As we have noted before, most of the time this has few consequences. Philosophical narratives, however, take us further away from rather than closer to the understandings of AI they promise to deliver. What is more, as we argued above, people act on the basis of the misunderstandings formal analytic conceptions of AI have helped promulgate. The conflation of 'Artificial Intelligence' with what we normally mean by the words 'artificial' or 'intelligence' (particularly the latter) is an error in the Rylean sense—we *assume* the link with what those words ordinarily refer to (i.e., human stuff) on the basis that the words are the same. But that leads us to talk in an ungrounded way—the problem of artificial intelligence is largely an artificial *problem*, which we've become 'bewitched' by (Wittgenstein 1953: §109) because of flawed reasoning that pitches the relationships between the terms 'artificial' and 'real'/'human' and the term 'intelligence'[9] as applied in both cases as ultimately comparable. There's no reason whatsoever to assume this. We don't *need* to do it. Indeed, proceeding on this basis leaves us more in the dark than when we started. Think of Wittgenstein's account of the ways in which we routinely categorise things as 'games':

> Consider ... the activities that we call 'games'. I mean board-games, card-games, ball-games, athletic games, and so on. What is common to them all?—Don't say: 'They *must* have something in common, or they would not be called 'games"—but *look and see* whether there is anything common to all. ... Are they all '*entertaining*'? Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. (Wittgenstein 1953: §66)

Here, Wittgenstein reminds us there are *lots* of things we call games and we mostly unproblematically know how to distinguish games from what aren't games (or quickly learn to do so). But there is no set of logical properties that constitute

---

[8] Our understanding of computational innovations is, for instance, poorer than our understanding of internal combustion engines, inasmuch as if our car were to break down whilst driving we might at the very least feel comfortable lifting the bonnet to see where the smoke is coming from. Contrast this with what happens when our computer technologies break down, where we may find ourselves thoroughly perplexed when confronted with the source code; unable to even begin to read it, even in the few cases where it is accessible to us.

[9] With the term 'intelligence' carrying its own baggage from a history of 'scientific' misuse as Mike Lynch has noted (personal correspondence).

a final checklist of conditions that must be satisfied if something is to be accurately called a game. Hence, treating the term 'game' as a formal one identifying a neatly delineated kind is a mistake—we shouldn't allow ourselves to be dazzled by it. However, there is little doubt that philosophers have been dazzled by AI, or rather, by their version of AI. That bedazzlement has spread and, as a result, we are all tempted to think in misguided ways about AI—primarily because we don't actually know what AI is. We thus get scared about the wrong things, put research money in the wrong places, spend time grappling with the wrong problems, come up with the wrong solutions, decide on the wrong courses of action, and more. We are also prone to blaming AI when 'it' goes wrong or 'rogue', when our efforts might be better repaid by thinking about how well or badly these technologies *have been made* to fit with the human world and our practices within it (Dreyfus 1992). A car may well be a salvation under certain circumstances (as a means of transporting badly-designed AI gadgets to the rubbish tip for instance) but few if any would think *it alone* has a capacity to deliver someone *to* salvation. It seems strange then, and perhaps belies a collective lack of understanding of such technologies, that we often attribute powers of salvation (or damnation) to computers running lines of code and get annoyed when *they* don't deliver. There are, as Brooks (1987) reminds us, 'no silver bullet[s]' to the problems of designing software capable of responding to increasingly complex contexts and practices of use and, upon reflection, we can see there is little reason to expect that there would be. With this in mind, we want to turn now to our examples—our 'perspicuous representations'—and what they have to teach us more specifically, first about the conceptual traps we have led ourselves into around AI but, second, also about how we might avoid them.

PERSPICUOUS REPRESENTATIONS OF AI

Following Ryle, Wittgenstein, Winch and Sharrock, our starting point here is the situated contexts of AI-relevant activities. We can approach such contexts in a number of ways. We can, for example, examine cases of AI as it is interwoven with and so implicated in human activity as a way of testing what it makes sense to say about AI and what it does not. Our approach is of this kind and our examples are thus less empirical studies than materials to think with, 'aids to a sluggish [sociological] imagination' (Garfinkel 1967; Mills 2000).

Tay: the Microsoft Twitter chatbot

Tay was a Twitter chatbot built by Microsoft; a computer program that leveraged a proprietary artificial intelligence algorithm to conversationally tweet with other Twitter users, and from those conversations, progressively 'learn' how to produce natural sounding conversations (i.e. engage in Twitter interactions in ways which

might evaluated as 'passing the Turing Test'). Tay was released on the 23rd March 2016, as a Twitter account/user who others could tweet at, thereby motivating conversational responses from Tay. Each conversation provided Tay with new data, on the basis of which the rules governing how Tay tweeted and the topics she[10] tweeted about were progressively adapted.

However, the design of the bot was problematic, inasmuch as it carried an assumption that all Twitter users who used Tay would do so benignly (i.e., that they would talk to Tay with the propriety and observe standards expected in conversations between humans). Anyone who has ever used the internet may already see the problem here. As Tay became more widely known, malicious users began to coordinate efforts to manipulate Tay's algorithm by force-feeding it hateful content (cf. Perez 2016). Recurrent themes were statements of extreme racism, misogyny and support for genocide often associated with the so-called 'alt-right' of Twitter. 'Hitler did nothing wrong!' and 'I fucking hate feminists and they should all die and burn in hell,' for example, are just two of the horrendous Tay outputs that displayed Tay's growing alignment with 'alt-right' sentiments. As an algorithm that uncritically and non-selectively 'learned' from its conversations with others on Twitter—whoever they were and whatever they might talk to Tay about—Tay's design and deployment had, in Microsoft's own terms, insufficiently 'anticipate[d] malicious intent that conflicts with our principles and values' and had subsequently been the subject of a 'coordinated attack by a subset of people' who 'exploited a vulnerability in Tay' (Lee 2016). The effect was that over the course of just 16 hours, Tay had produced nearly 100,000 tweets of an increasingly hateful and offensive character culminating in Microsoft's decision to shut down the service seemingly permanently to avoid further reputational damage.[11]

For the purposes of using the Tay affair as a way to provoke thinking around how we (mis)understand artificial intelligence technologies 'in the wild', there are several points worth making here. All of these highlight the unproblematic ways in which the distinction between artificial intelligence as a worked with technology (i.e., as a set of displayed its* we can interact with well enough for our practical purposes) and artificial intelligence as a philosophical concept (i.e., an account that attempts to turn the displayed its* of AI into a fixed formal category on which to base generalised philosophical claims about the 'true' nature of AI) was attended to in the malicious users' interactions with Tay.

Users' trolling of Tay's algorithm was premised on the exploitation of a simple vulnerability (which itself was the result of bad design on Microsoft's part)—it did

[10] There is a history to the tendency to attribute femininity to chatbots built to serve—Tay, Alexa, Cortana, Siri, etc—which is worthy of note (cf. Brooker 2019) but outside of the scope of the present chapter except as anthropomorphising grist to our argumentative mill.

[11] Though Microsoft suggested that Tay may return to Twitter in the future (Lee, 2016), to date this has not happened, nor has there been any further information to suggest that a re-tooled Tay is in the research pipeline.

not take very long or very much experimentation for users to see that Tay was a system for storing words/topics that would be used in future conversational exchanges with it. As it was the (human) user who initiated a topic of conversation with Tay, this gave those (human) users control over the inputting of keywords which would eventually re-emerge as outputs from Tay's account. Moreover, the displayed results of this experimentation—the its*—were completely public, affording users rapid and robust insights into how *the algorithm* operated—phrases that could be seen to be inputted into Tay earlier on could be traced as they made their way back into the public domain via Twitter's open platform. That the algorithm had no filter on content that would otherwise be identified as obviously inappropriate was clear from the ways in which coordinated efforts to suggest inputs became consolidated in due course as valid outputs.

In a similar vein, malicious users also quickly latched onto Tay's 'repeat after me' function, where a user could request that Tay repeat any statement verbatim, thereby adding it to her own vocabulary and bank of conversational contexts. In this way, malicious users could place topics and conversational forms *directly* into Tay, for the purposes of their being stored and recalled later. The phrase 'repeat after me' has very limited usage in human-to-human communication,[12] and its proliferation in the context of tweeting-to-Tay perhaps signifies that those deploying the phrase were doing something other than 'having a conversation' with her; indeed, their activity might be better described as 'gaming the algorithm'. Embedded in all of this is the idea that users evidently attended to the sense in which Tay's algorithm was probabilistic—the ability to experiment and *see* that common inputs were more likely to be reproduced as outputs was a clear demonstration that an effective way to manipulate the algorithm was to *coordinate* inputs using the public interface to maximise the likelihood of their future status as outputs. Tay's production of nearly 100,000 tweets itself means that nearly 100,000 requests were made of it, with the abiding topics of conversation—racism, misogyny and genocide—highlighting what those requests comprised.

Even when we attend to just these core mechanics of misappropriation, it is evident that users were not responding to Tay as if she were a genuine human co-conversationalist with a Twitter account—the relationship between Tay and these malicious users was not an interpersonal one but one of methodical experimentation in the direction of outputs by inputs, which suggests a treatment of AI's its* in ways that brush up against Microsoft's expectation that the bot would be engaged with as if it were akin to a human conversant (i.e., as an 'it' that would speak to the possibility of a machine acquiring and displaying a capacity to pass as human). The reminder this case serves us with is best understood via a

[12] We might imagine 'repeat after me' being used as a common prefix to utterances in a language classroom, speech therapy or singing lessons, where it may be important to practice and assess how subsequent phrases sound at least in the initial phases of a training or a treatment, but it has limited use beyond such settings.

Wittgensteinian (1953), Sacksian (1963) and Austinian (1962) treatment of these tweets not as propositions about the world but as movements in a game within it; i.e., by looking to see how a tweet 'does things with words' (Austin, 1962); in this case, how they were used in the gaming of an algorithm. And it is that distinction which makes specifically perspicuous the difference between Microsoft's anthropomorphic AI-as-'it' and the 'everyday' (though malicious) users' understanding that the AI-as-it*. The latter treats AI as fundamentally algorithmic (which is also to say, non-human), so that it can accordingly be made (even forced) to display at least some of its inner workings to render them available for experimentation and manipulation.

AI at/as work

Our second example shows the interplay of various stakeholders in trying to use AI in a commercial setting; in a company which develops artificial intelligence systems for business-relevant data analytics. In the work-episode we will focus on, AI is being utilised within an application to select relevant news articles for business users. More specifically, the purpose of the application was to scrape the world-wide-web for articles that referred to key business topics using keyword searches in real-time; this would then get passed to a machine-learning algorithm that would in principle reject articles that contained those keywords but were not relevant to business users (e.g., if the keywords were used in non-relevant and/or unanticipated contexts), and accept the ones that had the potential to inform business insights. The promise of the application was to keep the end-user more informed than their competition on topics of their choosing to ensure that they could make decisions better and quicker, avoid risks and grasp opportunities.

The building of the application crossed various practical domains: from business and management (to identify the correct semantic content), Development Operations or DevOps (as an approach to building software in ways that align with business objectives), to the more user-oriented requirements for functionality in the tool, data science (as the practical activity which sets the methodological norms for deriving insights from such a tool), and the explication of the tool in more general-purpose algorithmic but also commercial (i.e., marketing) terms. Different kinds of work task were being undertaken across those domains, and the ways in which those work activities converged on the algorithm were relevantly describable in different ways as a result. Rather than present an overview of all those activities and tease out the nature of those differences in terms of the several distinct but organisationally related practices they were embedded in, however, we shall examine a more limited set of engagements. The excerpts below were selected to help us focus on and contrast the development and assessment of the functioning of a natural language processing (NLP) algorithm from a data science perspective, with end-user's experience of putting the tool to work. More

specifically, our examination of the particular work episode in question will look at one aspect of how two members of the data science team, A and B, used (machine-driven) feedback to try to improve the performance of the AI system and the concepts they used to do this, as evidenced through messages sent via their internal (text-based) communications channels where such discussions are routinely held. While they had been satisfied prior to this that performance across specified parameters was adequate and thus that the algorithm was in good shape, end users' engagement with the displayed its* of the system led them to a rather different assessment—that they needed it to perform better still—leveraged via their vernacular competencies, 'what any business investor knows'. Their work was a response to those assessments.

This is a setting where AI technologies feature in and as routine elements of workplace activity. By following the work of the data scientists in building, adapting, and maintaining the algorithm, the picture of AI as some mysterious entity which is animated and made sentient rapidly loses whatever sense it may be thought to have. For the people directly involved in creating these algorithms, the work is more akin to the decidedly ordinary and non-mysterious activities of coding and statistics, which themselves comprise a set of well-defined and uncontroversial techniques. Moreover, in their practical uses of AI, this work team were perfectly capable of maintaining a separation between AI-as-'it' and AI-as-it*, in that the learning attributed to the machine *was* evaluated against the learning that humans can do but in a way that was practical and tangible rather than esoteric and intangible. In order to demonstrate this, we will work through the excerpts below which detail (text) conversations between the two data scientists, A and B, using an internal communication tool within the organisation. The conversation centred on an effort to increase the accuracy of the algorithm in order to make the results more relevant to the end-users whose feedback had occasioned this return to the algorithm. These efforts consisted of two parts. Firstly, producing a 'better' training data set, that more accurately and reliably identified positive and negative cases. Secondly, by using different methods to build in new features to form part of the algorithm.

The initial statement below by A is an expression of concern that the tool is creating problematic information from an end-user perspective. The machine learning algorithm is a supervised algorithm which means that the algorithm is trained to find patterns in the data on a known dataset. In this case, the data consist of sentences which contain keywords that refer to the business topic of interest and have the correct meaning that is required *and* those that do not. From there, the algorithm should be able to use this information to *learn* how to (mechanically) discriminate between relevant and non-relevant applications of the keywords, sorting them into different categories, in such a way as to derive a technique that can be applied to future examples. A's concern here is that the training data set may have some problematic data in it (and thus needs to be cleaned) that

is causing the algorithm to learn *incorrectly*. The other concern is that the type of algorithm they are using is not good at making discriminations on this type of data.

A: Hi—just finished training all of the data! Helped a little bit but we are only moving a few % points to the better—I wonder if we need to start looking at some different techniques? Do you want to send me a 'cleansed' training set over for me to look at and send back to you?

B replies to this with:

B: That's what I was planning on doing. Currently the program creates three outputs: noSUP_<filename>.csv, SUP_noRISK_<filename>.csv, and SUP_RISK_<filename>.csv. The 'cleansed' data is the final (since it matches and the keywords), but it might be worth running your trained eye over the other two to see if there is anything that is being scrubbed that shouldn't be.

Obviously I have all three for both the positive and negative training sets. Should I drop all 6 files here, or is there a better way?

Incidentally, we went from 3774 unique positive inputs to 2087. We went from 4743 unique negative inputs to 977.

Finally, I was having a think about perhaps using some sort of scoring function for how strongly positive or negative something is. I am not sure how the nlp [Natural Language Processing]/ml [Machine Learning] will handle it (but I'll find out a bit more today), but it means that those borderline examples will have a place to go.

I don't see why this would change anything on the user end (as in, we could try to come up with a way that works without changing the thumbs up/thumbs down interface)—but yeah, we can discuss that later.

While these excerpts involve technical shoptalk, they are followable. B agrees with A's strategy of cleaning the data, and furthermore suggests a way of scoring which would help the Natural Language Processing (NLP) algorithm, as an applied area of artificial intelligence, determine border-line cases, noting in passing

that the tool is already making more refined discriminations (with lower numbers potentially representing fewer false positives and negatives and thus tighter and more useful categorisation).

The exchange continues:

A: Hi, sounds awesome, yep send them over and I will have a look. Like the scoring idea. The current NLP is all over the place at the moment ...

B: Just an update: I have made a few changes to the ML code. Better feature selection pushes the positive recall on the old, incorrect training sets up to 0.65. Application of these methods to our cleaned data sets gives much better metrics (everything above 0.8), but that may be a consequence of our highly correlated training points. I have added a random predictor for comparison, and confusion matrix output for all ML predictors. I have also fixed the average prediction metrics, which isn't useful, to a predictor that uses a vote amongst the other predictors to make its decision. What we really need is to create some big training sets, which ties into the next point.

New Data: 538 training points. 83% positive. Old Data: 841 training points, 80% positive

5 fold cross validation:

New Data - 0.9646840148698885, 0.9823393997307039, 0.9756043956043955, 0.8798684210526316, 0.9084967320261438

Old Data - 0.8585017835909632, 0.894396594409083, 0.9351742919389979, 0.6813087633087633, 0.540530303030303

Values=[Accuracy, positive precision, positive recall, negative precision, negative recall]

In the last message B refers to the steps they have taken to increase the accuracy of the algorithm. Firstly, better 'feature' selection (a 'feature' being a measurable property of the phenomenon that can be added to the data such that the algorithm can use it to potentially make better predictions), secondly, using the new cleaned dataset (as expressed in the numerical results comparing the 'new' and 'old' data), and thirdly, creating a random predictor for comparison (hence able to estimate the performance of the algorithm better). Following this B evaluates the algorithm

with reference to both the old and new data by presenting a sequence of numbers between 0 and 1 (where a value of 0 is the lowest/worst and 1 is the highest/best). The old and new data give five such values each, and the key to their meaning is expressed at the end of the message 'Values=[Accuracy, positive precision, positive recall, negative precision, negative recall]'. B also qualifies these results by supplying additional contextualisation including, for example, 'training points' and their composition (for example, '83% positive' refers to the data being made up of 83% positive examples). B is also concerned to express the methods by which they arrived at this series of statistics ('5 fold cross validation').

What is evident from these excerpts is a sense that the language which A and B use to fence around the issue of improving and evaluating the performance of the algorithms (via the methods they have used to make the algorithm more 'intelligent') is quite divorced from what we ordinarily understand to be involved in the process of 'becoming intelligent' or being 'taught things'. As this brief work episode shows, the language used to create and refine the algorithms that make up AI's primary mechanism differ in pronounced ways from ordinary talk of learning or development. For instance, though the practitioners use the term 'training' to refer to both a dataset that provides a means of improving and evaluating of an algorithm *as well as* the skills of a human in evaluating the dataset's capacity to do so (A's 'trained eye', for instance), examining the uses of language in context shows that there is no conflation of the two different usages of the word to signify the same single process. Rather, what comes across here is that when we approach AI in the mundane context of work—as a product, a technique—we might better describe its 'intelligence' as being a mechanical matter of engineering 'performance management'. The notion of 'performance' involved is centred on statistical analysis, and the mechanisms to improve 'performance' include refining the training dataset, adding features to the algorithm, or changing the type of algorithm being used. And the results *of* that performance, and so the functioning of the tool, are made available *in* the displayed results, the its*. It requires no particular expertise to see that the values given by 'New Data' supersede those given by 'Old Data', other than the ability to recognise which numbers are higher than others.

Described in these terms, and seen as an occasioned practice, a response to end-users comments on the effectiveness of the system, none of these engagements with AI would lead us to treat the status of the technology as 'intelligent' or 'autonomous' except in strictly bounded ways (i.e., does its development and application bring the results closer to a value of 1 than before). In this context, the meaning of terms such as 'intelligence' and 'learning' are to be found in their locally intelligible uses (i.e., as attributes of AI-as-it*), nor does the practical production of a usable AI tool focus require those involved to grapple with epistemological and ontological questions or commit to particular metaphysical positions to resolve

them. Considerations of that kind are simply not important to getting the AI to work in and for the practical purposes at hand.[13]

Back to the source: code

Our final perspicuous representation is designed to bring out the sensibilities that surround the language we use to talk about AI, to probe if and/or where it may make sense to talk about AI anthropomorphically, and in what contexts. To do this, we offer a DIY *reductio ad absurdum*, where we boil down some of the core components of AI—that an AI is an autonomous or unsupervised system that can progressively refine its own outputs[14]—and work through what these may look like in code. Though there are many languages, libraries and modules that afford a programmer access to sophisticated artificial-intelligence-related technologies (e.g., machine learning), we avoid these here, on the grounds that they tidy away the relevant algorithmic goings-on from users. This is typically done for ease of use, though here it would obscure what might be taken from reading the code on-screen. Instead, we aim to write what we see as the simplest forms of code that may be describable in relation to a small selection of AI-relevant concepts (namely, 'autonomy' and 'learning') and display them as fully as we can. Our aim is to test just where those descriptions might or might not apply.

Taking the idea of autonomy and un-supervised operations first, here is a program (written in the Python programming language) capable of producing an output that is practically impossible for a human to generate or predict in advance:[15]

```
01. from random import randint
02. print(randint(0,100))
```

[13] We might put this another way: even when we turn to the sites where AI is being practically realised, we do not find AI as a thing-in-itself. Instead, hands-on engagements with AI cannot be extricated from 'what-we're-all-doing-here-now' as part of local courses of practical action and reasoning and the practical projects they are constitutive of. Strip away the practical context, therefore, and talk of AI loses all specificity and meaning, as AI is to be found in its practical involvements.

[14] We recognise that there can be more nuance to a definition of AI than this (but often not much, see Shanker 1987). However, for the purposes of a perspicuous representation (as opposed to a comprehensive empirical accounting), we argue that this rough-and-ready definition captures the sense of AI as the field focused on the production of computationally 'intelligent agents' (cf. Kasabov 1998; Russell and Norvig 2003), which is one particularly pervasive way of defining AI within the computer science community at least.

[15] This code was written specifically in/for Python version 3.6.3., though barring any unforeseeable changes to the core language and syntax of future versions, the authors expect that this code will work for any past or future Python 3 installation, should a reader wish to try it for themselves. The numbers shown at the beginnings of each line of code are simply for reference—line numbers to be referred to in the ensuing description—and do not make up functional elements of the code.

Line-by-line, this code first imports a technique for generating random integers from the Python library that deals with randomisation ('from random import randint'), and then uses that technique to print out a randomly-generated number between 0 and 100 ('print(randint(0,100))'). Or at least that is one way of describing it. Another (more anthropomorphising) way might be to say that this is an algorithm that is capable of choosing to say something—a number between 0 and 100—in a way that has an extremely high probability of confounding even the best guesses of a human who might try to predict what the outcome will be. Of course, this *isn't* artificial intelligence in the standard sense. But, asking *why* this is not artificial intelligence—why we might prefer the more technical description (the it*) over the more open-ended cognitively-oriented one (the 'it')—helps us think through how we understand AI to operate.

In the code excerpt above, the result is something that is produced by the computer, in such a way that there is a limited (1 in 100) chance of predicting beforehand what it will do. However, though the computer selects the result, the parameters around which a valid result might be produced (and which also set the range of sensible predictions a human might advance) are baked into the code by whoever wrote it. Is this autonomy (as one of the hallmarks of an artificially intelligent system)? Is the computer doing anything unsupervised? It could be, but only if we *choose* to talk about the code in that way, something which itself involves conceding that whatever unsupervised autonomy we might ascribe is at best a shorthand that absolves us of learning how to read Python code. And for those of us uninterested in learning Python code, such an anthropomorphic description—a notion of the code as an 'it' of a sort—may perfectly serve our needs. For instance, it allows us to deal with the outputs in isolation and engage in characterising them. For philosophers who seek to then take the code's displayed its* and formalise them into fixed objects of epistemological and ontological investigation—AI-in-itself—a closer look at the code would suggest that this would be a significant conceptual misstep.

It is possible to throw increasingly complex code at this particular problem, and if we are sufficiently able to read it (or at least understand a line-by-line redescription of it), we might note that complexity doesn't itself add any grounds to claim that any piece of code is approximating any of the cognitively-oriented terms we may informally use to talk about it (cf. Shanker 1987: 635). For instance, as noted above, another claimed hallmark of artificial intelligence is the capacity for a program to progressively refine its own outputs—in anthropomorphic terms, to change its own behaviour on the basis of learned information. To build code that does *that* work, we will need to add further layers to the two lines of code above as follows:

```
01. from random import randint
02.
```

```
03. learned_numbers = []
04. total_iterations = 0
05.
06. def rng():
07.     global learned_numbers
08.     number = randint(0,1000)
09.     if number in learned_numbers:
10.         pass
11.     else:
12.         learned_numbers.append(number)
13.
14. while len(learned_numbers) < 1000:
15.     total_iterations = total_iterations + 1
16.     rng()
```

Line-by-line, this code imports the same random integer generation technique as described above ('from random import randint'), and establishes placeholder variables into which information can be stored ('learned_numbers = []' and 'total_iterations = 0', both of these will be referred back to later). Following on from this, we define a function (lines 6 to 12 inclusive) for randomly generating numbers ('def rng():') which we allow access to 'learned_numbers' as the place where we are going to store numbers that are learned by the algorithm ('global learned_numbers'). We then generate a random integer between 0 and 1000 and store that in a variable called 'number' ('number = randint(0,1000)'). Following on from this, we have part of a function for learning to avoid certain numbers—an if/else condition that stipulates if our recently-generated-random-number is already stored in 'learned_numbers' we pass over it, and if it is *not* in that list of 'learned_numbers', to put it in there. On the back of this function, we then establish a 'while loop' (lines 14 to 16 inclusive) which stipulates that until the program has learned all the numbers from 0 to 1000 ('while len(learned_numbers) < 1000:', or, in other words, while the list of learned numbers still has less than 1000 entries), two things will happen. First, a value of 1 will be added to the previously-assigned total_iterations variable ('total_iterations = total_iterations + 1', which effectively counts how many times the program has tried a randomly-generated number regardless of whether that number has been learned already or not). Second, the random number generator function will be invoked ('rng()') to randomly generate a number in the ways described for code lines 6 to 12 above.

To re-describe the program as a whole, it consists of a set of instructions that ask Python to go through as many iterations as it takes to populate a list of numbers from 0 to 1000, with each number being randomly generated each next time the process runs and is checked against a 'learned numbers' list as it develops. The difference between the value of 'total_iterations' (which will be different each time

the program is run, depending on the random numbers generated thus far) and the length of the 'learned_numbers' list (which will always end up at 1000) is the amount of times the algorithm has tried and rejected a random number (on the basis that it has already learned to avoid it). As the program learns more numbers to avoid, it becomes increasingly less likely that the random numbers generated are still unknown/un-encountered—on our most recent running of the program, this happened 5893 times (as given by the sum 'total_iterations—len(learned_numbers)').

We have already established that it does not make sense to say that code such as this is autonomous. However, do the later additions and layers of complexity make it sensible to claim that this program can 'learn' to avoid already 'known' numbers as it iterates through a range of numbers from 0 to 1000? Again, it comes down to whether we *choose* to describe it in one way or another—both forms of description may (and will) apply depending on the contexts in which those descriptions are situated and the ends to which they are being put. That is, it is the practical context, not the technology, that grounds the drawing of this distinction. Proponents of AI as an 'it', however, routinely run considerations of context and technology together. Such elisions may generate radical and provocative claims, but they have little relevance to understandings of the technologies they are making those claims on behalf of.

Until algorithms are connected up with other working mechanisms and our practical affairs, and until we have enough of a handle on them to describe them sensibly within those local contexts, they may be anything; or nothing (cf. Wittgenstein 1953: §12). How we make the distinction is key. Concentrating on the capacity of our second code excerpt to 'learn' anything, we might note that what this code does can just as well be described in other ways—we might say that it 'stores' results for use in 'conditional logic statements' that gradually add iterations until a pre-defined outcome is reached (i.e., 'the length of a list reaches 1000'). As with the first code excerpt, however, paying closer attention to the code itself, and describing it in its own locally intelligible terms, affords a way of steering clear of the kinds of anthropomorphic conflations of displayed its* with a free-standing thing-in-itself, AI-as-'it', that have generated misconceptions about AI since it was first explored as a technical possibility in the 1930s, and which bedevil the field still.

DISCUSSION: CHOOSING THE LANGUAGE OF AI

All three of our perspicuous representations bring to the fore that idea that a lack of attention to the different ways we may choose to speak of AI—the language we may variously invoke in its description and to what effect—accounts for the continued philosophical confusion evident around it. Having chosen to use a set of descriptors which map AI technologies onto human cognition in various ways,

analytic philosophers have too readily run with the supposed 'problems' in the wrong direction, failing to first examine whether these problems, and the assumptions they reflect, are sensible starting points for investigation. Our article, grounded as it is in the practice-oriented, contextual and thus ethnomethodological approach advocated by Sharrock and colleagues over the course of many years (see Anderson and Sharrock 1979; Button, Coulter, Lee and Sharrock 1995; Sharrock and Anderson 2011; and Sharrock and Button 1999, as a woefully incomplete selection) would indicate that they are not.[16] What the perspicuous representations above have been intended to show, and what Sharrock and colleagues have advised all along, is that attending to how AI is talked about and worked with in mundane everyday settings teaches us why AI does not have to be treated as in need of being philosophically grounded at all—in Wittgensteinian (1953) terms, 'its' status as a philosophical problem of any kind dissolves when we examine our multiform involvements with these technologies in these ways.

Doing the work of dissolution here mandates a focus on the 'occasioned character' (Sharrock and Ikeya 2000: 275) of our practical engagements with AI, that is, the local contexts and applications within which its sense is determined and rendered available to participants and analysts alike. In the case of Tay, for instance, we see understandings of how its algorithm worked clearly on display in and through the ways malicious users manipulated inputs to promulgate hateful and offensive content. This was possible given the local context of Twitter as a public space with particular conversational/interactional conventions and Tay's encoded affordances as a chat-bot, two features of the setting those seeking to game the algorithm quickly figured out how to exploit when the occasion arose. Regarding the suggested improvements and negotiations around the effectiveness of AI as a tool for decision-making in financial investments and business management, A and B's engagement with the algorithm was occasioned by a need to improve its workings so as to generate more relevant output from an end-user perspective and shows us how alternate interpretations of those technologies and their capabilities could be aligned. This was done by using the device of assigning numerical values between 0 and 1 to 'old' and 'new' versions of the algorithm that enabled those versions to be compared against one another through the measurement of performance. The tricky part, of course, is that these assessments were not just in the hands of A and B but also had to satisfy the requirements of customers who evaluate the performance of the machine against other practical criteria,

[16] Those who know Wes Sharrock well will hear echoes of one of his favourite apocryphal stories here: A, wandering the streets of an unknown city on the way to find their lodgings, becomes lost. A spies B, walking down the street towards them. Hailing B as they pass, A asks B how they can get to their lodgings from the street they're both on. Responding to A, B answers as follows; 'Well; I wouldn't start from *here*'. Too much contemporary work on AI has failed to heed the implicit advice: if we are genuinely seeking understanding of how AI technologies work, we need to be careful about where we begin from conceptually. Not all 'heres' are equal (again, cf. Shanker 1987).

criteria their model of performance had to accommodate. Similarly, the code excerpts presented above represent an occasioned response to the treatment of anthropomorphic conceptions of AI in fixed, formal analytic terms as picking out essentialised features of AI systems; a critical aim we stated at the outset.

In attending to the occasioned character of such practical engagements with AI, we have sought to demonstrate the ways in which those engagements exhibit and account for their own intelligibility (cf. Sharrock and Anderson, 2011). Though it may make sense on occasion to use anthropomorphic language to describe AI, what it is and what it does (we can hopefully drop quotation marks and asterisks by this stage and return to ordinary language), we can see from the work involved that there are plentiful grounds to reject the wholesale import of those terms into philosophical argument. The putatively 'technical' repurposing of ordinary language neglects the practical, contextual and occasioned character of the use of those terms and the wider forms of activity they acquire meaning within, and wilfully disregards the point that what makes sense in particular contexts will rapidly descend into nonsense when treated as divorceable from any context. Rather than facilitating understandings of AI as a feature of our lives, such decontextualising approaches will always block attempts to produce useful descriptions of those practical engagements with specific aspects of the technology in *any* particular setting we may be trying to make sense of.

CONCLUDING REMARKS

We mentioned at the outset that contemporary AI, 'New AI', as it is often called has sought to disentangle itself from cognitivism and the dream of real artificial intelligence by adopting an engineering rather than a philosophical perspective on developing AIs; namely, how can we use algorithms, processing power and code to develop new technologies that perform useful tasks for us? However, while this shift in orientation back to AI in something closer to the way it was first conceived in the 1930s, where it was recognised it would be a logico-grammatical error to talk of machine's as 'intelligent', is welcome, it is also important to recognise that it remains incomplete. All too often in fields like machine learning, the idea that algorithms are learning is taken at face value and discussions arise as to how 'they' can be made to 'learn' more effectively. Rather than the picture of a digital mind, we have here a picture of digital artefacts doing mind-ed or mind-ful things. Similarly, while machines may no longer be intelligences, their operations are routinely described as exhibiting 'intelligence' and developers ask how AI in general can be made to exhibit 'intelligence' more reliably. As we noted, this installs new agentic ghosts in the algorithmic machine, ones given an existence by virtue of the qualities of the tasks they perform. For this reason, we felt it was important to return to the problems associated with the New AI as a means of dissolving the philosophical rather than technical or engineering problems which lead people to

conceive things in this way. To come at last to the epigram at the beginning of the article—Wittgenstein's remarks on the *human* character of these machines—we wanted to emphasise that it makes no sense to talk of AI in any way outside of particular contexts, particular projects and the particular practical purposes AI technologies are designed—well or badly—to fulfil. Once again, algorithms are part of our lives, not separate from them, and they gain their practical relevance and consequentiality by virtue of where and how they feature within them. What is more, while we do interact with algorithms, we do not do so in vacuo. Instead, wherever we find an algorithm, we do not need to look far to find the set of human beings who developed it and the sets of human beings putting it to work, or attempting to do so. Algorithms are working artefacts which must be conceived, produced and maintained; they are not self-generating. The human and the machine are thus deeply interlinked and when we assess algorithms, we are frequently evaluating how well or badly we think the designers have done in producing them in a particular way.

Our aim in the article has thus been to demonstrate that if we fail to place AI in its practical contexts (in the ways advocated throughout), we start to see ghosts in our machines. We end up with no way of accounting for the things we want to account for, i.e., the workings of AI, and are seduced into ascribing those workings to categories that make them more rather than less opaque. Moreover, in the process we link one set of black boxes (e.g., code and programs) to others (e.g., 'learning') in ways that become progressively difficult to unpack. However, if we *do* take up the practically-oriented approach Sharrock and colleagues have consistently recommended, we will see that there are no ghosts, and indeed, there are barely any machines either.[17] As the body of work in which Wes Sharrock has been a central figure suggests, the entirety of the issue hinges on the language with which we choose to talk about AI (i.e., the sense we wish to make of it). As much of the philosophy of mind has done already, we could choose to conflate cognition with the anthropomorphic terminology we apply to computational technologies as a neat shorthand and thereby make thorny philosophical problems out of proposed points of parity and begin to doubt the distinctions between humans and machines. However, if we *do* choose to do that, the work presented here suggests that we would also need to recognise that we are thereby placed in the realm of science fiction rather than describing how AI features in the world as it is.

[17] At least no machines that we are not already comfortable with: calculators, statistics, computer chips, etc. Admittedly such technologies may be sophisticated, but we may still hesitate to equate sophistication and intelligence directly.

REFERENCES

Anderson, Digby C., and W. W. Sharrock. 1979. 'Biasing the News: Technical Issues in "Media Studies"'. *Sociology* 13 (3): 367–85.

Austin, J. L. 1962. *How to Do Things with Words*. Cambridge, MA: Harvard University Press.

Brockless, Adrian. 2019. 'Thought, Consciousness, Brains and Machines'. *Philosophy Now* 130: 16–19. Available at: https://philosophynow.org/issues/130/Thought_Consciousness_Brains_and_Machines (Accessed: 16 May 2019)

Brooker, Phillip. 2019. 'My Unexpectedly Militant Bots: A Case for Programming-as-Social-Science'. *The Sociological Review*. Available at: https://doi.org/10.1177/0038026119840988 (Accessed: 16 May 2019).

Brooks Jr., Frederick P. 1987 'No Silver Bullet: Essence and Accidents of Software Engineering'. *Computer* 20 (4): 10–19.

Burrell, Jenna. 2016. 'How the Machine "thinks": Understanding Opacity in Machine Learning Algorithms'. *Big Data and Society* 3 (1): 1–12.

Button, Graham, Jeff Coulter, John R. E. Lee, and Wes Sharrock. 1995. *Computers, Minds and Conduct*. Cambridge: Polity Press.

Chollet, Francois, and J. J. Allaire. 2018. *Deep Learning with R*. Shelter Island, NY: Manning Publications.

Coulter, Jeff. 1979. *The Social Construction of Mind*. Basingstoke: Macmillan.

Coulter, Jeff, and Wes Sharrock. 2007. *Brain, Mind and Human Behavior in Contemporary Cognitive Science*. Lewiston, NY: Edwin Mellen Press

Dennett, Daniel. 1987. *The Intentional Stance*. Cambridge, MA: Bradford Books.

Dreyfus, Hubert L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. London: MIT Press.

Fodor, Jerry A. 1981. 'On the Impossibility of Acquiring "More Powerful" Structures: Fixation of Belief and Concept Acquisition'. In *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky*, edited by M. Piattelli-Palmarini. Cambridge, MA: Harvard University Press.

Garfinkel, Harold. 1967. *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice Hall.

Garfinkel, Harold. 1990. 'The Curious Seriousness of Profession Sociology'. In *Les Formes de la Conversation*, edited by B. Conein, M. de Fornel and L. Quéré, 69–78. Paris: CNET.

Garfinkel, Harold. 2002. *Ethnomethodology's Program: Working Out Durkheim's Aphorism*. Lanham, MD: Rowman and Littlefield.

Kasabov, Nikola. 1998. 'Introduction: Hybrid Intelligence Adaptive Systems'. *International Journal of Intelligent Systems* 6: 453–54.

Lee, Peter. 2016. 'Learning from Tay's Introduction', *Official Microsoft Blog*. Available at: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (Accessed: 16 May 2019).

Mackenzie, Adrian. 2015. 'The Production of Prediction: What Does Machine Learning Want?'. *European Journal of Culture Studies* 18 (4–5): 429–45.

Mills, C. Wright. 2000. *The Sociological Imagination*. Oxford: Oxford University Press.

Perez, Sarah. 2016. *Microsoft Silences its New A.I. Bot Tay, After Twitter Users Teach it Racism [Updated]*. Available at: https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/ (accessed: 24 May 2018).

Runciman, David. 2018. 'Diary: AI'. *London Review of Books* 40 (2): 38–39.

Russell, Stuart J., and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach*. 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Ryle, Gilbert. [1949] 2000. *The Concept of Mind*. London: Penguin.

Sacks, Harvey. 1963. 'Sociological Description'. *Berkeley Journal of Sociology* 8: 1–16.

Sacks, Harvey. 1992. *Lectures on Conversation*. Oxford: Blackwell.

Schutz, Alfred. 1962. *Collected Papers*, vol. 1: *The Problem of Social Reality*. The Hague: Martinus Nijhoff.

Shanker, Stuart. 1987. 'Wittgenstein Versus Turing on the Nature of Church's Thesis'. *Notre Dame Journal of Formal Logic* 28 (4): 615–49.

Sharrock, Wes. 1989. 'Ethnomethodology'. *British Journal of Sociology* 40 (4): 657–77.

Sharrock, Wes, and Bob Anderson. 2011. 'Discovering a Practical Impossibility: The Internal Configuration of a Problem in Mathematical Reasoning'. *Ethnographic Studies* 12: 47–58.

Sharrock, Wes, and Graham Button. 1999. 'Do the Right Thing! Rule Finitism, Rule Scepticism and Rule Following'. *Human Studies* 22 (2–4): 193–210.

Sharrock, Wes, and Nozomi Ikeya. 2000. 'Instructional Matter: Readable Properties of an Introductory Text in Matrix Algebra'. In *Local Educational Order*, edited by S. Hester and D. Francis, 271–88. Amsterdam: John Benjamins Publishing Company.

Sharrock, Wes, and Jeff Coulter. 2004. 'ToM: A Critical Commentary'. *Theory and Psychology* 14 (5): 579–600.

Sharrock, Wes, and Jeff Coulter. 2009. '"Theory of Mind": A Critical Commentary Continued'. In *Against Theory of Mind*, edited by I. Leudar and A. Costall, 56–88. London: Palgrave Macmillan.

Silver, David, et al. 2016. 'Mastering the Game of Go with Deep Neural Networks and Tree Search'. *Nature* 529: 484–89.

Silver, David, et al. 2017. 'Mastering the Game of Go Without Human Knowledge'. *Nature* 550: 354–59.

Stich, Stephen. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press.

Stilgoe, Jack. 2018. 'Machine Learning, Social Learning and the Governance of Self-Driving Cars'. *Social Studies of Science* 48 (1): 25–56.

Suchman, Lucy. 2019. 'Apparatuses of Recognition'. Jackman Humanities Institute Marquee Lecture, University of Toronto, 22 April.

Turing, Alan M. 1950. 'Computing Machinery and Intelligence'. *Mind* 59 (236): 433–60.

Winch, Peter. 1958. *The Idea of a Social Science and its Relation to Philosophy*. London: Routledge & Kegan Paul.

Wittgenstein, Ludwig. 1956. *Remarks on the Foundations of Mathematics*. Oxford: Basil Blackwell.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell.

Wittgenstein, Ludwig. 1959. *Remarks on Frazer's Golden Bough*. Atlantic Highlands, NJ: Humanities Press.