Bert Olivier

Prof. Bert Olivier, Department of Philosophy, University of the Free State, South Africa; E-mail: olivierG1@ufs.ac.za; bertzaza@yahoo.co.uk

DOI: http://dx.doi.org/10.18820/24150479/aa49i1.1

ISSN: 0587-2405

e-ISSN: 2415-0479

Acta Academica • 2017 49(1): 2-21

© UV/UFS

Artificial Intelligence (AI) and being human: What is the difference?

First submission: 28 March 2017
Acceptance: 4 May 2017

This paper begins by focusing on the recent work of David Gelernter on artificial intelligence (AI), in which he argues against 'computationalism' – that conception of the mind which restricts it to functions of abstract reasoning and calculation. Such a notion of the human mind, he argues, is overly narrow, because the 'tides of mind' cover a larger and more variegated 'spectrum' than computationalism allows. The argument of Hubert Dreyfus is examined, that the AI research community concentrate its efforts on replacing its cognitivist approach with a Heideggerian one, a recognition that AI research cannot ignore the 'embeddedness' of human intelligence in a world, nor its 'embodiment'. However, Gelernter and Dreyfus do not go far enough in their critique of AI research: what is truly human is not just a certain kind of *intelligence*; it is the capacity for 'care' and desire in the face of mortality, which no machine can simulate.

Keywords: Artificial Intelligence (AI), care, computationalism, embodiment, mind

1. Introduction: Is human 'intelligence' synonymous with AI?

Usually, when artificial intelligence (AI) is compared to human intelligence, it is done in predominantly,

if not exclusively, 'quantitative' terms; that is, along lines which 'measure' AI's capacity for abstract 'reasoning', such as calculation. This is intuitively the case, judging by numerous conversations ordinary people have about AI in various forms, most often that of the computer, where it is usually discussed in terms of its powerful calculating abilities. An instance of this was where the specially built AI machine 'Deep Blue' defeated the reigning world chess champion, Gary Kasparov, who played in several matches against it in 1996 and 1997, to mixed reactions. (Deep Blue versus Gary Kasparov; Gelernter 2016: xiii). The point is that, whether one's response was shock or delight, it related to Deep Blue's AI capacity to defeat a human being seen as the embodiment of supreme human intellectual ability, where the latter was limited to abstract, calculative performance. At a more reflective level, this general tendency, to restrict one's assessment of AI to its calculating power, was also acknowledged by Sherry Turkle in her early work, *Computers and the Human Spirit*, although she was more interested in its function as an "evocative object, an object that fascinates, disturbs equanimity, and precipitates thought" (1984: 19).

This view of AI, predominantly in terms of its calculative capacity, again emerges in the recent book, *The Tides of Mind: Uncovering the Spectrum of Consciousness* (2016), by computer scientist David Gelernter of Yale University in the United States, where he intimates that most contemporary computer scientists and 'philosophers of mind' seem to believe that 'mind' (which includes their conception of AI) is confined to the abstract, logical, high-focus functions of so-called 'rational' thinking (2016: xii-xiv). In answer to the question, "Why should philosophy of mind be obsessed with digital computers?", he points out that (2016: xii):

> There are three explanations, all related. One centers on computers as a test-bed for mind theories. Another focuses on computing as a powerful, simple way to describe or blueprint events in time: *processes* – that is, organized actions. The last explanation, a theory called computationalism, asserts that brains *are* computers, and the mind is just software that runs on the brain. This would be awfully neat if it turned out to be true.

In what follows, it becomes clear that Gelernter is convinced that it is not the case, however – something that contrasts sharply with the view of Georg Schwarz, expressed in 1990, where he optimistically remarked (Schwarz 1990: 2): "Computationalism assumes that cognitive systems compute functions; the existence of non-computable functions would serve to refute it only if these functions could be shown to be constitutive for cognition. So far, this has not happened." Arguably, it *has* happened with the publication of Gelernter's book, insofar as it is precisely 'non-computable [cognitive]

functions' of the human mind that he concentrates on. While computationalism attempts to understand the 'parallel' functioning of the mind and the brain, on the assumption that 'computing' or calculation, according to a set of 'algorithmic' steps or rules, is key to gaining insight into this, Gelernter focuses on the non-computable cognitive functions of the mind. In short, he singles out all those functions that computers, or AI in its present guise (and I shall argue in *any* guise, for reasons outlined below), cannot subsume under the aegis of computation and computationalism. At the present stage of its development AI is known as 'weak AI' – that is, AI that depends on human programming for performing a variety of computational tasks – in contrast with what AI researchers associated with the programme known as 'strong AI' are working towards, namely AI that would be the equivalent of humans as conscious, thinking beings capable of everything humans can do, and more (Armstrong 2017). Alongside the strong AI research programme there is the burgeoning field of 'connectionism', which Medler (1998: 63) described, more than 17 years ago already, as a theory of "information processing" (within the broader field of cognitive science) which deviates from those systems that utilise (algorithmic) rules hierarchically for the manipulation of symbols. Instead, the cognitive models of connectionism mimic the neurophysiology of the human brain by focusing on the development of "parallel processing", as it occurs in the human brain. Whatever the differences between these two research programmes might be, however, they share, in my view, the reductionist principle of equating the mind with a kind of software and the brain with hardware, which is – particularly in light of Gelernter's work – untenable. Why?

First, Gelernter (2016: xviii-xix) comments on the unsurprising fact that philosophers of mind have been struck by the resemblance between computers and the functioning of the human brain, where the mind is compared to software and the brain to hardware. This should surprise no one; after all, human beings designed the computers to perform *some* of the functions their minds usually carried out, but *not all*. Needless to add: this does not logically imply that minds *are* computers; for one thing, humans ('embodied minds') have life-stories, while computers do not. (More on this below.) The first of three explanations for philosophers of mind noting the similarity between computers and human brains, referred to by Gelernter, has to do with "theoretical AI", namely, the opportunity provided by AI to ascertain whether theories about the nature and functioning of the mind are correct, such as when a theory of language acquisition on the part of AI (but by implication also on the part of 'the human mind') is put to the test in a "working program" (2016: xii-xiii). While theoretical AI therefore concerns rule-following behaviour, Gelernter (2016: xiii) points out, "applied AI" focuses on problems that minds (of the human as well as the AI variety) can only solve by

using their intelligence, instead of following a set of embedded rules. The case of Deep Blue scoring an AI victory over Gary Kasparov in chess, referred to earlier, illustrates the successful outcome of such an 'applied AI' programme.

The second explanation for philosophers of mind being preoccupied with the way that digital computers operate, according to Gelernter (2016: xiii-xiv), is that it provides them with the opportunity and the "framework" to understand how "*processes*" or "actions-in-time", in the shape of a series of prescribed steps (which involve condensing and varying the meaning of lists of instructions), work. Importantly, this relates to "following a prearranged set of rules", in other words, to an algorithmic function of the mind, which, with the theme of this paper in mind, is qualitatively different from the question of the putative capability, on the part of AI, to make ethical judgements. Anyone who imputes to moral or ethical decisions by humans or, by analogy, on the part of an AI, an underlying algorithmic function, would unavoidably do so by thinking in a reductionist manner, that is, by reducing human beings to machines, with no freedom of will or choice. I realise that the latter is a position defended by some thinkers, but – as Kant (1960: 30) showed in the late 18th Century – unless one presupposes freedom of the will in human beings, it makes no sense to talk about ethical or moral behaviour, because, without freedom of the will, all ostensibly 'ethical' (or, for that matter, 'unethical') human actions would be predetermined by some other 'law' or force.

It is the third reason for the overwhelming interest of philosophers of mind in digital computers that really interests Gelernter – the assertion that human brains (which use software called 'minds') *are* indeed computers, albeit embodied differently. This position is what gave rise to "computationalism" (2016: xviii) along the line of reasoning that goes more or less as follows: computers compute, but because computing is a form of thinking, namely, clear, rational thinking, computers can be called 'thinking machines'. And because the 'intelligent' performance of such 'rational', thinking tasks can be carried out by computers, and thinking is something done by minds, it seemed to follow that studying the theory, performance and structure of digital computers was tantamount to studying the mind. Hence the philosophical field of 'computationalism'. In the rest of this complex book, and in opposition to the majority of other computer scientists, Gelernter makes out a case for the difference between "brain" and "mind", elaborating on the distinctive character of the mental activity known as 'free association', as opposed to focused, conscious mental activity, and on the crucial contribution of fantasy and dreaming to creative thinking – in short, all those mental capacities in which human beings routinely engage without a second thought. The irony should be clear, that, in a world where there is a growing propensity, particularly among philosophers of mind and computer scientists, to use something conceived and constructed by human beings,

namely the computer, *reductively* as a model to comprehend what it is to be human, a philosophically-minded computer scientist such as Gelernter sets out to demonstrate that there is a fundamental difference between AI in the guise of the computer and being human, or more precisely, the human mind in all its diverse 'tides'.

What do these 'tides' entail (most of which are ignored by computationalism)? It should already be apparent that Gelernter diverges from what computer scientists who subscribe to futurologist Raymond Kurzweil's techno-optimism (2006: 39-46) believe, namely, that humanity is on the cusp of a technological development that will, in a mere few decades from now, yield to the advent of the so-called 'Singularity', when (according to Kurzweil) AI will immeasurably surpass human intelligence, and pave the way for humans to merge with machines. Gelernter, on the other hand, instead of prostrating himself before the god of technology, reminds his readers forcefully about the dimensions ('tides') of the human mind of which AI has not even scratched the surface, as it were. Accordingly, as a philosophically-minded computer scientist, Gelernter delves into the work of literary giants such as Shakespeare, Tolstoy, Proust and J M Coetzee, as well as that figure who (more than anyone else, perhaps) changed the way human beings think about themselves, to wit, the inventor of psychoanalysis as a novel discipline, Sigmund Freud. Only by scrupulously examining the fruits of mental accomplishments across a whole 'spectrum' of activities can one hope to gauge the mind's true 'depth', after all. In other words, instead of restricting one's investigation to the high-focus, logical functions of so-called 'rational' thinking, as most computer scientists and philosophers of mind do, Gelernter considers the mind across what he calls a "spectrum". The latter stretches from "high focus" mental activities, such as strongly self-aware reflection, through "medium" engagements such as experience-oriented thinking (including daydreaming accompanied by emotion) to "low focus" functions like "drifting" thought, with concomitant emotions flourishing, and dreaming (2016: 3; 241-246). Furthermore, the different roles of memory are significant here. According to Gelernter, at the "high focus" niveau of the mental spectrum, memory is employed in a disciplined manner, while at the medium-focus level it "ranges freely" and when the low-focus level is attained, memory "takes off on its own".

One might wonder what the point of such a 'psychological' investigation by someone in computer science might be. As far as one can gather, what Gelernter hopes to achieve by delineating this "spectrum" is to demonstrate as graphically and persuasively as possible that the human "mind" is multi-dimensional, instead of being the uni-dimensional faculty that computationalism makes it out to be. This is why he demonstrates that the mind is characterised by different "tides", *all of which* belong to it irreducibly, instead of only the one that is located at the

level of "high focus". It is the latter that conventional AI research has claimed as its exclusive province, to the impoverishment of (one's conception of) the mind. For Gelernter such research, on the part of the 'mind sciences' in general, concentrates solely on the high-focus level of mental functions, in the misguided belief that this *alone* is what 'mind' is, in both its AI and human embodiments (2016: xi-xix). Moreover, such a uni-dimensional conception of the mind cannot possibly do justice, in his view, to the nature of *creativity* that, in the final analysis, marks an insurmountable difference between AI, on the one hand, and creative human intelligence, on the other.

Regarding the question of AI and ethical consciousness, the following passage articulates Gelernter's view, if it is kept in mind that computalionalism equates AI and (the human) mind (2016: 22): "The scientist explains the origin of the universe with a logical argument. The religious believer tells a story... Only the logical argument has predictive power. Only the story has normative moral content. Only a fool would pronounce one superior." The domain of logic coincides with that of AI and the human mind, as conceived by computationalism, which, as implied by these observations, is devoid of moral significance. The domain of stories, or narrative, on the other hand, is redolent with moral significance, and comprises a domain wholly different from, if not alien to, that of cold logic. This, I take it, is tantamount to saying that, while AI – built as it is on the basis of the computationalist model of the mind – excels at the level of logic, morality, or action that is ethically meaningful, is incommensurate with its capabilities. However, Gelernter does not explicitly thematise this, which is arguably a shortcoming in his attempt to demonstrate the incompatibility between AI and human 'intelligence' in the broadest sense of the term, which, I would argue, includes the capacities of making ethical (and aesthetic) judgements (Olivier 2008). After all, although an ethical judgement of, say, a person's actions being corrupt, is not in itself a 'cognitive judgement' that yields knowledge 'for its own sake', or for theoretical purposes, it does presuppose 'knowledge' of an ethical, moral, or 'practical' (as derived from 'praxis') kind. In terms of Gelernter's notion of the 'spectrum' of mental activities, one might say that moral or ethical judgements and actions are rooted in a level that he does not explicate, although it is imbricated with some of the mental 'tides' he distinguishes. In fact, I am constantly surprised by the fact that thinkers who are at pains to discern the irreconcilable differences between AI and human beings mostly seem to focus on the cognitive functions of the mind alone – even if it is done with great sensitivity for the entire 'spectrum' of such functions, as in Gelernter's case – to the exclusion, surprisingly, of human moral or ethical attributes that one might insist are wholly outside of the reach of AI as it is currently conceived, and perhaps even in principle. This is the case even where some of the most knowledgeable philosophers are concerned, and also in a

certain sense when some philosophers seem to grasp the importance of the issue of the ethical, as I shall show below.

First, however, an elaboration on what was earlier referred to in passing is called for. As everyone should know, at least on the basis of a little reflection, computational tasks are only *some* (and certainly not all) of those carried out by human minds, as Gelernter demonstrates at length. These tasks also include aesthetic judgements, in addition to the ethical, noted above, and perhaps the most striking feature of human beings – which could be called a function of 'embodied' human minds – namely the fact, noted by Jacques Lacan (2007: 206-215; Olivier 2005), among others, that every human subject has her or his own personal narrative (even if an analyst may be needed to reconstruct its continuity by gaining access to the 'censored chapter' of their lives, their unconscious – something I cannot pursue here at length, despite its importance as a differentiating factor regarding humans and AI; see Olivier 2005a). In fact, perhaps the most fundamental thing shared by humans the world over, regardless of language and culture, is that all individuals have a personal narrative or life-story (embedded in more encompassing cultural narratives such as communal or national myths), which has nothing to do with the abstract, computational tasks associated with AI in the form of computers.

Jean-Francois Lyotard gives a striking account of the centrality of narrative in human life where he elaborates on what he calls the "pragmatics of narrative knowledge" (Lyotard 1984: 18-23). Knowledge, Lyotard reminds one, is not synonymous with science or learning, but includes ideas of "know-how", "knowing how to live" and "how to listen" (1984: 18). These ideas are connected with "traditional knowledge", the most important form of which is "narrative" (1984: 19). To clarify what is at stake in narrative, Lyotard further discusses "popular stories" (including myths), the diverse "language games" that narratives accommodate (such as denotative statements, deontic or duty-oriented statements, interrogative and evaluative ones), and the rules for the "transmission of narratives" (usually pertaining to cultures of orality, but also the generational telling of stories in families) (1984: 20-21). Then there are narratives' "effect on time", such as the "rhythm" of ritual performances of tribal tales, but also of fairy tales ('Once upon a time...happily ever after') and nursery rhymes, which are always "contemporaneous with the act of recitation" (1984: 21-22), and, finally, the "authority" that narratives have in relation to traditional knowledge (1984: 22-23). The importance of this for the present theme is that it emphasises the difference between humans and AI in terms of "narrative knowledge", which is unthinkable for entities (AI) that have no grasp of temporal succession as story, narrative or 'personal history'. To clarify this further, elsewhere Lyotard (1991: 15) elaborates from another, related angle on the differences between AI

and human 'thinking', where he states that the "...main objection [to equating human intelligence with AI] concerns the very principle of these intelligences. Our disappointment in these organs of 'bodiless thought' comes from the fact that they operate on binary logic..." In a manner that resonates with much of Gelernter's argument about the "tides of [human] mind", he continues as follows (1991: 15):

> But as Dreyfus [discussed below] argues, human thought doesn't think in a binary mode. It doesn't work with units of information (bits), but with intuitive, hypothetical configurations. It accepts imprecise, ambiguous data that don't seem to be selected according to preestablished codes or readability. It doesn't neglect side effects or marginal aspects of a situation. It isn't just focused, but lateral too. Human thought can distinguish the important from the unimportant without doing exhaustive inventories of data and without testing the importance of data with respect to the goal pursued by a series of trials and errors...This picture inevitably recalls the description Kant gave of a thought process he called reflective judgement: a mode of thought not guided by rules for determining data, but showing itself as possibly capable of developing such rules afterwards on the basis of results obtained 'reflexively'.

One could add that Lyotard also leans on the work of Maurice Merleau-Ponty (whom he proceeds to refer to in the chapter quoted from) in this argument, which is unnecessary to discuss at length here (see Olivier 2002). The point here is that some of the same features of human thinking that enable individuals to listen to (or tell) someone's story – pre-eminently among them time, or temporal immersion – also operate in human thinking more generally, in contradistinction to the attributes of AI as indicated.

2. Dreyfus on what AI research lacks

Gelernter is not the only one who is dismissive about the current direction of AI research. Hubert Dreyfus, a Heidegger and Merleau-Ponty scholar, known for his polemic against those AI researchers who claimed that they would solve the problem of human intelligence where philosophy had failed for more than 2 000 years, has drawn on his knowledge of these thinkers to criticise the direction of such research. His critique of 'intellectualist' AI research is initially that (ironically, given their apparent contempt for philosophy) it was too exclusively modelled on rationalist philosophy's conception of concepts as rules, which it proceeded to formalise in AI programmes predicated on the assumption that intelligence was a function of making inferences from internal symbolic systems, whether these

were in the minds of people or digital computer memories (Dreyfus 2007: 1). As a phenomenology scholar, it was clear to him that what AI research needed was a substantial injection of Heideggerian understanding of intelligence as embodied, and as responding to, as well as learning from, one's environment, an insight that, when the opportunity presented itself, he communicated to AI researchers at MIT and elsewhere, with a noticeable effect on the direction in which some of them subsequently took their work (2007: 1-14). But even the most comprehending and best-intentioned among them could not quite get it right to turn the AI research programme into a 'genuine' Heideggerian project. The one who evidently came closest to translating Heidegger's fundamental ontology of *Dasein's* (the human being's) involvement with the concrete world through 'readiness-to-hand' (which corresponds with things that are 'ready-to-hand'), Phil Agre, acknowledged Dreyfus's Heideggerian influence on him in passages like this one (quoted in Dreyfus 2007: 10):

> I believe that people are intimately involved in the world around them and that the epistemological isolation that Descartes took for granted is untenable. This position has been argued at great length by philosophers such as Heidegger and Merleau-Ponty; I wish to argue it technologically.

To grasp what is at stake here, a brief clarification of 'readiness-to-hand' in Heidegger's work is necessary. When human beings enter into a relationship with things such as hammers and screwdrivers as 'ready-to-hand', it differs from approaching things as 'present-at-hand', The former refers to a pragmatic relationship of 'use' with what Heidegger calls "equipment" which can be 'manipulated' – as when one hammers with the instrument by that name, and the kind of being of such 'equipmental' things is "readiness-to-hand". When things are 'present-at-hand' one is not in a pragmatic relationship with them, but in one that merely acknowledges that they are 'there'. Most of the time this would be the case with things like trees and mountains, but even things approached as 'ready-to-hand' can, under certain circumstances, be perceived as 'present-at-hand', for example when a hammer breaks and its pragmatic functionality is undermined. Then, until it has been repaired, it is merely 'present-at-hand' (Heidegger 1978: 95-107).

Agre's promising 'translation' of Heidegger's ready-to-hand things in the world, which are not separated from human beings in a cognitivist sense (where one can only 'reach' them by somehow moving from 'internal' representations to 'external' objects), but are instead directly accessible to embodied humans who are always already in the world and in touch with the things around them, floundered in the end. He was on the right track, however,

in Dreyfus's estimation, with the idea that 'facts' in the world had to be replaced by "possibilities for action that require appropriate responses from the agent" (2007: 10), which were programmed using "deictic [pointing] representations" that corresponded with what he thought of, by analogy with Heidegger's ready-to-hand, as "deictic intentionality". Dreyfus comments approvingly (2007: 10) that Agre had understood Heidegger's ready-to-hand correctly, not as a "*what* but a *for-what*". So what went wrong? In a nutshell, Agre and his colleagues were working with a misleading conception of what it means to be human in relation to the world we inhabit, and moreover, they were restricting their work to an idea of the cognitive relations between humans and the world. More specifically, from Dreyfus's account (2007: 11-12), it is clear that Agre, like all the other AI researchers before him, made the fatal mistake, with his "deictic representations", of *objectifying* an act that, in the human lifeworld, is one of immediacy – of responding to the 'what-for' of something like a door or hammer, for example – by linking a function as well as its "situational relevance" to a programmed rule that would determine an agent's response. What Agre was trying to replicate in his AI programme was a human situation of direct involvement with the world, instead of one that is 'distanced' from the world through representations. Ironically, this repeated the mistake of 17th-Century empiricists like John Locke, who believed that one knows the world on the basis of the experience of things as represented in the mind, which had the effect of enclosing the knowing human agent within the circle of their own representations. It took the work of figures such as phenomenologists Edmund Husserl and Maurice Merleau-Ponty (both of whom Dreyfus refers to; see for instance pp. 3, 28), as well as Heidegger, to demonstrate that what humans know is not representations of the world but the world itself. The other AI researchers whose work is discussed by Dreyfus in the rest of the paper similarly fail to provide the formal AI counterparts of embedded, embodied human being-in-the-world. If AI research wishes to replicate a genuinely human relationship with the world, it has to find ways to make their programming of 'AI agents' *more* Heideggerian, as Dreyfus persuasively argues. In his 'Conclusion' Dreyfus remarks (2007: 30):

> It would be satisfying if we could now conclude that, with the help of Merleau-Ponty… we can fix what is wrong with current allegedly Heideggerian AI by making it more Heideggerian. There is, however, a big remaining problem. Merleau-Ponty's… account of how we directly pick up significance and improve our sensitivity to relevance depends on our responding to what is significant for *us* given our needs, body size, ways of moving, and so forth, not to mention our personal and cultural self-interpretation. If we can't make our brain model responsive to

> the *significance* in the environment *as it shows up specifically for human beings*, the project of developing an embedded and embodied Heideggerian AI can't get off the ground.
>
> Thus, to program Heideggerian AI, we would not only need a model of the brain functioning underlying coupled coping… but we would also need—and here's the rub—a model of *our particular way of being embedded and embodied* such that what we experience is significant for us in the particular way that it is. That is, we would have to include in our program a model of a body very much like ours with our needs, desires, pleasures, pains, ways of moving, cultural background, etc.

From the preceding discussion it should be clear that, as a critic has reminded me, many AI researchers, by engaging in the kind of work that Agre and others (discussed by Dreyfus) have done, acknowledge that algorithmic-computational AI models represent but a tiny part of the full spectrum of human cognition (as highlighted by Gelernter, too). Among these researchers one might single out the name of Andy Clark, who similarly attempted to demonstrate, in Heideggerian fashion, that intelligent beings have to be able to interact with an environment – as the provocative title of his book (1997), *Being There: Putting Brain, Body, and World Together Again*, testifies.

However, although Dreyfus touches upon the most crucial aspect of the irreducible difference between humans and AI in the last sentence in this excerpt (where he mentions desires and cultural background), embodiment is not neutral regarding ethical action. As Heidegger (1975) has shown, human desire is linked to the earth through the human body, and arguably the ethical presupposes desire, because, as Kant (2002: 31-32) so clearly indicated in his second *Critique*, and Lacan in his Seventh Seminar (*The Ethics of Psychoanalysis*; 1997, pp. 311-325), the problem of ethical or moral choice confronts one in the field of human desire. What AI research needs is not simply, as Dreyfus has argued, to become 'more Heideggerian' in its programming of AI by factoring in an equivalent of 'being-in-the-world' in terms of readiness-to-hand, which would presumably enable a suitably 'embodied' AI to respond to things that are ready-to-hand by learning how to *use* them appropriately. What is required is something far more radical, and probably, I would guess, unattainable, namely building the counterpart of what Heidegger calls 'concern', which is a variety of 'care' (with its implication of desire), into AI. I have argued along these lines before (Olivier 2008) in light of precisely such a challenge posed to AI research by two science fiction films. These are Spielberg's *AI* and Proyas's *I, Robot*, where, thematically speaking, the former challenges the AI research community to realise their objective of producing a human simulacrum that is capable of *caring* or *desiring* (incarnated in the

robotic boy, David), and the latter pitches this challenge at the level of a robotic being capable of *ethical* consciousness, as manifested in its capacity for *guilt*. Here I would like to follow a different approach, however, by focusing in a more sustained manner on something implicated by Dreyfus's critical engagement with AI theorists, but not developed by him in the paper referred to, namely Heidegger's contention, that the core structure of *Dasein* (human being) is *care*.

3. Care as the distinctive ontological trait of human beings

My reason for drawing attention to these considerations is to emphasise that even a Heidegger scholar like Dreyfus does not go far enough in his critique of mainstream AI research programmes. By restricting his critique to the broadly 'cognitive' level of the ready-to-hand, he omits what it is ultimately a manifestation of, namely *concern*, which (for Heidegger), is an ontologically fundamental structural attribute of *Dasein* (human being), namely (most fundamentally) *care*, of which *concern* is an expression (Heidegger 1978: 237). When one uses a hammer, which is ready-to-hand, in this way actualising one's immediate involvement in the world, without the supposed intervention of mental representations (ideas, etc.), it is a manifestation of one's 'concern' for the things in the world (the chair one is fixing, as part of the furnishings of one's home, for example), which, in turn, indexes one's core ontological structure as *care* (Heidegger 1978: 237):

> Because Being-in-the-world is essentially care, Being-alongside the ready-to-hand could be taken in our previous analyses as *concern*, and Being with the Dasein-with of Others as we encounter it within-the-world could be taken as *solicitude*. Being-alongside something is concern, because it is defined as a way of Being-in by its basic structure – care. Care does not characterize just existentiality, let us say, as detached from facticity and falling; on the contrary, it embraces the unity of these ways in which Being may be characterized. So neither does 'care' stand primarily and exclusively for an isolated attitude of the 'I' towards itself. If one were to construct the expression 'care for oneself' ['Selbstsorge'], following the analogy of 'concern' [Besorgen] and 'solicitude' [Fürsorge], this would be a tautology. 'Care' cannot stand for some special attitude towards the Self; for the Self has already been characterized ontologically by 'Being-ahead-of-itself', a characteristic in which the other two items in the structure of care – Being-already-in...and Being-alongside... – have been *jointly posited* [*mitgesetzt*].

The excerpt, above, should be read in conjunction with this one (Heidegger 1978: 236; see also 67-68) "Dasein is an entity for which, in its Being, that Being is an

issue". To unpack the full meaning and implications of these excerpts from *Being and Time* is impossible in a mere article – the extent of the secondary literature on this difficult, but important book (which anyone can easily ascertain) is testimony to this. However, one may summarise the meaning of Heidegger's claim, that human beings are essentially characterised by 'care', as follows, taking into consideration the other important concepts referred to in these excerpts as well.

For Heidegger the distinctive ontological mark of *Dasein* (human being) is precisely that her or his being – the very fact that they *are*, or exist – is not merely, mutely, accepted; it is 'an issue' for every human being in the sense that she or he 'does something about it, and with it'. One can either live one's 'factical' existence by 'falling', that is, living in accordance with the dictates of convention and fashion, or one can elaborate on one's being as an 'issue' by designing a 'project' for oneself, which may take one far afield from conventional preoccupations – although one cannot avoid convention altogether; one's singularising 'project' invariably takes root within the realm of convention, although it surpasses it to the degree that it belongs to an individual *Dasein*. In both cases – 'falling' into convention, or following one's own, distinctive 'project' – the way one lives is a manifestation of 'care' (whether it is 'caring' what your 'friends' think of you on fashionable social media sites, or 'caring' about the outcome of your unique 'project', such as designing environmentally sustainable houses). This also means that one is always 'ahead-of-oneself', which here means that one projects the care-structure of one's being into the way one lives in the light of one's orientation to the future. What a first-year student does at university is explicable in terms of what future they envisage for themselves, even if it rests on what they did in the past. The aetiological emphasis lies in the future as far as all human actions are concerned – this is what Heidegger means by claiming that human 'selves' are typically 'ahead-of-themselves'. Moreover, just as humans always reveal their 'caring' in every future-oriented situation, they reveal it in their 'concern' with things (like furniture that has to be repaired) and in their 'solicitude' for the well-being of other people (which can be either affirmative or negative; whether I wish my friend well, or the person who is competing with me for promotion not well, both are manifestations of 'solicitude'). Notice here that 'care' is fundamental – it is not something over and above concern and solicitude – this is why Dreyfus's argument is implicitly, but should be *explicitly*, embedded in this more encompassing Heideggerian conception regarding human beings' ontological care-structure. All of this is further compatible with what was argued, earlier, that humans are distinct from AI insofar as every person has a life-story or narrative, embedded in a more encompassing cultural narrative; the future-orientedness of *Dasein*, combined with its 'project'-character – even if it

is not affirmatively 'taken up' – implies that *Dasein* has a history, and therefore a 'story'.

As a benevolent critic has reminded me, one should add that care as a fundamental ontological structure of *Dasein* which involves facticity, existentiality, and fallenness, is not restricted to these, but is further essentially related to human mortality, or what Heidegger (1978: 279; 289) calls "Being-towards-death" or "Being-towards-the-end". It is important to note that, although Heidegger (1978: 281) stresses the essential "Being with Others" of every *Dasein* or person – a relatedness to community that is distinctive about human beings, in contradistinction from AI – who can relate 'objectively' to other people's deaths, this is not possible regarding one's own death, nor can another person "represent" or stand in for anyone else when it comes to dying (Heidegger 1978: 284). Moreover, Heidegger stresses – significantly, regarding differences between human beings and AI – that "Death, in the widest sense, is a phenomenon of life" (1978: 290), which means that only living beings can die (even if there are significant differences, according to Heidegger, between the dying of human beings and the "perishing" of animals and plants; p. 284). Heidegger talks about the "existential-ontological structure of death" (1978: 293), pointing out that it must be understood in terms of "...the fundamental characteristics of Dasein's Being: existence, in the 'ahead-of-itself'; facticity, in the 'Being-already-in'; [and] falling, in the "Being-alongside". In other words, ontologically, "...*dying is grounded in care*" (1978: 296). What this means, in the final analysis, is that every human being has to face his or her own death as "...the possibility of the absolute impossibility of Dasein" (1978: 294) in existential terms (as opposed to a mere physiological or biological process of perishing; pp. 284, 291), and they can (either) do it in such a way that they yield to 'falling factically' into the 'inauthentic' sphere of what Heidegger calls "everydayness" (1978: 296-299), which has the effect of tempting one "...to cover up from oneself one's ownmost Being-towards-death" in various ways (p. 297), or – instead of such evasion or "tranquillization" (p. 298) in the face of death – one can accept the 'indefinite certainty' of one's death (p. 302), in accordance with one's temporal, future-oriented 'being-ahead-of-oneself', by 'anticipating' one's temporally unspecified yet certain death as your "*ownmost* possibility" (Heidegger 1978: 307), because "anticipation" coincides with the very kind of temporal being of *Dasein*, or being human. In fact, Heidegger argues, 'anticipation' of one's own death as something "non-relational" that cannot be "outstripped", "frees" oneself "*for* accepting this" (1978: 308). Furthermore, in the face of the "nothing" of death, *Dasein*'s being is essentially one of "anxiety", as opposed to "cowardly fear", which imparts to one a "freedom towards death" (1978: 310-311). Finally – skipping over Heidegger's lengthy analysis of all the aspects involved – one arrives at the insight

that a person (*Dasein*) who 'attests' to an 'authentic' existential relationship with their 'ownmost possibility of non-being', that is, with her or his being-towards-death, arrives at this through a specific response to the silent 'call of conscience', which is simultaneously the 'call of care'. The liberating response, or "choosing" in question is called "resoluteness" by Heidegger (1978: 314).

It is supremely doubtful whether any instance of AI is capable of this 'anticipation', let alone existential anxiety in the face of certain, though unspecified, death, because if it ceases to exist, for whatever reason, its cessation cannot possibly be synonymous with the multi-faceted death of a time-bound, inescapably mortal human being, as evocatively characterised by Heidegger, and succinctly reconstructed above. As already stated, *Dasein* is a being whose being is 'an issue for itself'; hence, its death is no exception in this regard. One might say that if, in the light of *Dasein's* temporality or future-directedness, it has a life-history, story or narrative, death – towards which every person adopts a specific existential relationship, either denying it in various ways, or 'resolutely' facing it as one's own 'possibility of non-being' – marks the end, or singular possibility of non-being, of every particular person's unique narrative.

In a later publication Heidegger (1975: 143-159) provides an encompassing axiological framework within which his earlier analysis of human being in terms of 'care' (including 'solicitude' and 'concern') can be situated. Briefly, here he outlines four principles ('earth, sky, mortals and divinities'), or fundamental values, called 'the fourfold', which comprise the 'horizon' against which human life makes sense, or gains significance. These four principles form a unity, and without *all* of them – or rather, what they stand for – functioning together, human cultural practices would, by implication, founder. This is because they encompass what it means to be human, as may be gauged from the following brief clarification. 'Earth' denotes, as the word suggests, the earth as the ground of human existence and embodiment, which means it also stands for human needs and desires. 'Sky', by contrast, is an indication, for Heidegger, of that inscrutable source of what humans are subject to, such as good and bad weather, but – as Karsten Harries (1997: 152-166), in his discussion of Heidegger's 'fourfold' observes – also of a limit that challenges humans to surpass or overcome it in their striving. 'Mortals' – probably the most significant of the 'fourfold' in the present context, given the impossibility that intelligent machines can 'die' – is a reminder that humans must all, unavoidably, die as humans, and that this temporally bounded life is the only opportunity that each person has to live meaningfully. The final principle, 'divinities', does not allude to a specific god, God, or gods, but to those largely inaccessible, but axiologically indispensable, sources of significance for every human being, whether these are the gods of specific religious traditions, or simply forces and powers that are regarded, or experienced, as imparting value

to one's life. It should be clear that these four principles correlate with the 'care-structure' of *Dasein* set out by Heidegger in *Being and Time*, and arguably, only human beings, whose lives are 'framed' by them, can either live meaningful – or, alternatively, meaningless – lives; no AI in the guise of a robot is capable of this. And it is my argument that the primary reason for this is that humans are mortal, as graphically indicated by the principle of 'mortals' in the fourfold; robots don't die, and don't have to find meaning in their lives. This is an irreducible difference between humans and AI.

Carol Gilligan's (1982) so-called 'ethics of care' – which she formulated to account for the differences between a typically masculine, rule-oriented ethical approach to the world, and a typically feminine, care-oriented approach – could be seen as one of the implications of Heidegger's characterisation of human beings' way of being-in-the-world, albeit one oriented towards women, although it does not exclude the possibility that men could adopt such an ethical stance. This is the direction in which AI research should go, but I have a nasty suspicion that it would be a dead-end street. After all, how do you programme desire, care, or concern, in the erotic, affective and ethical sense(s) of these terms, into AI in the form of a computer or a robot? Where Sherry Turkle (2010: 5-6) differentiates between AI and human beings, she articulates well what is at stake here, in a manner that resonates with Heidegger's understanding of human beings:

> I am a psychoanalytically trained psychologist. Both by temperament and profession, I place high value on relationships of intimacy and authenticity. Granting that an AI might develop its own origami of lovemaking positions, I am troubled by the idea of seeking intimacy with a machine that has no feelings, can have no feelings, and is really just a clever collection of 'as if' performances, behaving as if it cared, as if it understood us. Authenticity, for me, follows from the ability to put oneself in the place of another, to relate to the other because of a shared store of human experiences: we are born, have families, and know loss and the reality of death. A robot, however sophisticated, is patently out of this loop...What kinds of relationships with robots are possible, or ethical? What does it mean to love a robot?... A love relationship involves coming to savor the surprises and the rough patches of looking at the world from another's point of view, shaped by history, biology, trauma, and joy. Computers and robots do not have these experiences to share.

In addition to reinforcing the argument that AI, or robots, do not have the personal stories or narratives that every human being is privy to – the 'experiences' she invokes – what Turkle touches on here when she writes "behaving as if it cared"

resonates with Heidegger's emphasis on 'care', but also brings another perspective, that of a robotic being that has been programmed to respond to 'caring' or affectionate (human) touch by behaving, or more accurately, performing 'as if' it was responding in kind. One witnesses this kind of 'behaviour' on the part of AI in the guise of 'therapeutic' robotic pets that have been programmed to 'show affection' to certain ways of being touched (and not to others, like being handled roughly), Turkle (2010: 8-9) informs one, which provide comfort to, for instance, lonely, elderly people. She discusses an instance of this, where a seventy-two-year old lady's attachment to her robotic 'pet' (called a 'Paro') – a "sociable robot in the shape of a baby harp seal" (Turkle 2010: 8) – is apparent from the way she strokes and talks to it, while the Paro, as if it were a real, live pet, 'responds' by 'looking' at her and purring (because it has been programmed to react to being treated in such a gentle manner). Just how performatively convincing these 'as if' modes of behaviour, or, more accurately, programmed modes of performance, are, is apparent from the comfort that this elderly lady gets from her interaction with the robotic 'pet'. But there is a catch; in Turkle's words, referring to the lady in question (2010: 8): "In attempting to provide the comfort she believes it [the Paro] needs, she comforts herself". Through the artificial mediation of an 'intelligent' machine, the illusion is created that it is the machine that comforts her. This discussion occurs in the context of Turkle's account of AI's transition from calculating, 'thinking' machines (computers) to 'affectionate' machines in the form of robots. Whereas formerly computers were regarded as being somehow 'alive' because of these intelligent machines' demonstrable ability to 'think', more recently robotic beings' ontological status as 'living beings' has been measured by a different criterion, namely, whether they exhibit affection (Turkle 2010: 2, 26-32), with particularly children judging affirmatively in this regard.

The crucial question is, however, whether one can legitimately talk about 'affection' here, and it is clear from what Turkle writes in the excerpt, quoted above, that she does not believe this to be the case. Instead, one is faced with a programmed performance. Even more important: when children judge robotic AI to be 'somehow alive' on the basis of their performance, which *simulates* affection on the basis of pre-programmed, performative criteria, it should alert one to a telling contradiction. Robots can be, and are, pre-programmed to respond to human touch, *as if* they were alive, which they are not. This means that they do not die – something which the individuals who benefit from them therapeutically may find reassuring – or more accurately, they *cannot* die, even if they can be destroyed. Only beings with a personal history of joy and suffering, as evoked by Turkle in the earlier quotation, can die (either in agony, or peacefully, with resignation, or in protest about their finitude; whatever the case may be); AI, which has no inkling of such experiences, could never be privy to it, particularly in

light of the fact – testified to by Gelernter and Dreyfus, above – that AI research has not managed to overcome the problem of rising above a narrow, cognitivist programming model.

4. Conclusion

From the preceding discussion, the inadequacy of contemporary attempts to produce an AI that is a 'true' simulation of a human being should be apparent. I write 'a human being' advisedly, because the point of the discussion has been to show that even the description 'human intelligence' is wrong and misleading. 'Intelligence' is a multi-faceted word, and what it denotes cannot be reduced to 'logical' or 'calculative' functions, but pertains to what Heidegger called the 'ready-to-hand' as well (thoroughly demonstrated by Dreyfus), and beyond 'intelligence' the whole panoply of human experience made possible by humans' 'care-structure' – their defining capacity to experience the world in a wide variety of modes of caring – still beckons as the (in my view unattainable) goal of AI research. It is one thing to programme a robot to behave *as if* it cared, and quite another to programme it to being *able* to make judgements, rooted in an ontological care-structure (or in the 'fourfold'), similar to that of humans, about ethical issues for example, such as whether to support or condemn someone who rejects certain people on the basis of their cultural preferences.

Bibliography

Armstrong A (2017) Artificial intelligence – strong and weak. http://www.i-programmer.info/babbages-bag/297-artificial-intelligence.html (Accessed 23 May 2017).

Clark A (1997) *Being there: Putting brain, body, and world together again*. Cambridge, MA: The MIT Press.

Deep Blue versus Gary Kasparov. https://en.wikipedia.org/wiki/Deep_Blue_versus_Garry_Kasparov (accessed 7 March 2017).

Dreyfus H (2007) Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. http://leidlmair.at/doc/WhyHeideggerianAIFailed.pdf (accessed 9 March 2017).

Gelernter D (2016) *The tides of mind: Uncovering the spectrum of consciousness*. New York: Liveright Publishing Corporation.

Gilligan C (1982) *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Harries K (1997) *The ethical function of architecture*. Cambridge, MA: The MIT Press.

Heidegger M (1975) Building dwelling thinking. In: *Poetry, Language, Thought*. Trans. Hofstadter A. New York: Harper Colophon.

Heidegger M (1978) *Being and time*. Trans. Macquarrie J and Robinson E. Oxford: Basil Blackwell.

Lacan J (1997) *The seminar of Jacques Lacan – Book VII: The ethics of psychoanalysis 1959-1960*. Trans. Porter D. New York: W.W. Norton.

Lacan J (2007) *Écrits. The first complete edition in English*. Trans. Fink B. New York: W.W. Norton & Company.

Lyotard J-F (1984) *The postmodern condition: A report on knowledge*. Trans. Bennington G and Massumi B. Manchester: Manchester University Press.

Lyotard, J-F (1991) *The inhuman. Reflections on time*. Trans. Bennington G and Bowlby R. Cambridge: Polity Press.

Kant I (2002) *Critique of practical Reason*. Trans. Pluhar W. Indianapolis: Hackett Publishing Company, Inc.

Kant I (1960) *Religion within the limits of reason alone*. Trans. Greene TM and Hudson HH. New York: Harper Torchbooks.

Kurzweil R (2006) Reinventing humanity: The future of machine-human intelligence. *The Futurist* (March-April): 39-46. http://www.singularity.com/KurzweilFuturist.pdf (Accessed 15/07/2016).

Medler DA (1998) A brief history of connectionism. *Neural Computing Surveys* 1: 61-101. http://www.blutner.de/NeuralNets/Texts/Medler.pdf (accessed 23 May 2017).

Olivier B (2002) Body, thought, being-human and artificial intelligence: Merleau-Ponty and Lyotard. *South African Journal of Philosophy* 21(1): 44-62. https://doi.org/10.4314/sajpem.v21i1.31335

Olivier B (2005) Lacan and narrative identity: *The Piano Teacher*. In: *Word, (wo) man, world: Essays on literature. Festschrift* for Ina Gräbe. Oliphant AW and Roos H (eds) Pretoria: UNISA Press. Reprinted in Olivier B (2009) *Philosophy and psychoanalytic theory. Collected essays*. London and Frankfurt: Peter Lang Academic Publishers.

Olivier B (2005a) Lacan and the question of the psychotherapist's ethical orientation. *SA Journal of Psychology* 35(4): 657-683. Reprinted in Olivier B (2009) *Philosophy and psychoanalytic theory. Collected essays*. London and Frankfurt: Peter Lang Academic Publishers.

Olivier B (2008) When robots would really be human simulacra: Love and the ethical in Spielberg's *AI* and Proyas's *I, Robot*. *Film-Philosophy* 12 (2). http://www.film-philosophy.com/index.php/f-p/article/view/56/41

Schwarz G (1990) What is computationalism? http://90.146.8.18/en/archiv_files/19902/E1990b_107.pdf (accessed 23 May 2017).

Turkle S (2005) *Computers and the Human Spirit*. Cambridge, MA: The MIT Press.

Turkle S (2010) *Alone together: Why we expect more from technology and less from each other*. New York: Basic Books.