Is 'Intelligence' a Sufficient Criterion for Establishing Equivalence of AI with Being-human?

Bert Olivier
ORCID iD: https://orcid.org/0000-0002-3138-1948

Abstract

This paper thematises the difference(s) between Artificial Intelligence (AI) and being-human, with a view to answering the question, whether it is enough to compare AI and human beings in terms of (a very narrow conception of) intelligence. First the question of human distinctiveness is raised, and a number of human attributes are enumerated, such as the singularity of every human being, the individual's capacity for rebellion or revolt – as elaborated upon by Albert Camus and Julia Kristeva, and with historical as well as fictional examples of this. The significance of Kristeva's insight into the link between revolt and pleasure is also noted, before turning to David Gelernter's investigation into the true compass of the human mind, as compared to the 'computationalist' conception of mind, as it is used by AI research to measure the success of creating AI commensurate with human 'intelligence'. Finally, the issue of the capacity, on the part of AI such as robots, of mimicking human behaviour, for example deceiving, is examined with a view to demonstrating what AI research has to equal if it hopes to be successful in constructing a true android. This is explored through an interpretive analysis of the science fiction film, *Ex Machina* (Garland), which is carried out to demonstrate the true challenge to AI research, to produce a convincing human simulacrum.

Keywords: Artificial Intelligence (AI), being-human, singularity, rebellion, revolt, deception

Print ISSN 1023-1757; Electronic ISSN: 2519-5476; DOI https://doi.org/10.29086/2519-5476/2021/sp38a8

Introduction: AI, 'Rebellion' and the Distinctively Human

Recently I had a conversation with someone who uncritically believes that, in the future, Artificial Intelligence (AI) will resolve all 'problems' faced by humans. I put 'problems' in scare quotes because there is only one kind of 'problem' that AI can resolve, or 'solve', namely problems of a technical kind – such as those hypothetically posed in the science fiction narrative of Ridley Scott's *Blade Runner* (Scott 1982), which involve tasks in deep space requiring superhuman strength, and which the so-called 'replicants' are technically designed to carry out. The potential problem of these human simulacra rebelling against their makers – the way the fictional Cylons do in *Battlestar Galactica* (Moore 2004 - 2009) and the machines in the *Terminator* films (Cameron 1984; and 1991; Olivier 2002) – is also dealt with in technical terms in the *Blade Runner* narrative by giving the replicants a limited, built-in four-year lifespan, which, ironically, is precisely why some of the replicants rebel in the story. But even if fictional AI is depicted as rebelling against their human creators, this is arguably no more than an anthropomorphic attribute required by the narratives in question, and not what one encounters on the part of AI (in the guise of computers, for example) today. Moreover, the technical problems that AI is designed to tackle do not circumscribe, much less define, what human life distinctively entails. The fictional Mr Keating – teacher of English in Peter Weir's profound film, *Dead Poets Society* (Weir 1989; Olivier 2002a) – articulates this well when, after instructing them to tear out part of their poetry books (where poetry is erroneously reduced to a mere formula by the editor) he informs his class of eager-to-learn boys that he has 'a little secret' for them: 'We don't read and write poetry because it's cute', he asserts; 'We read and write poetry because we're members of the human race. And the human race is filled with passion!' He continues: 'Medicine, law, business and engineering; these are noble pursuits and necessary to sustain life; but poetry, beauty, romance, love – these are what we stay alive for ...'. Then he concludes by appealing, as it were, to their capacity for making imagination the source of giving direction and purpose to their lives – that is, to the *unique* life of each and every one of them. He emphasises the decisive importance of the fact that they are *there* and '... life exists; and identity'. Moreover, that 'the powerful play goes on', and they may 'contribute a verse'. Looking around at the boys' rapt faces, he directs a final question at (by implication each one of) them:

'What will your verse be?' In passing I should briefly note that, as an anonymous critic has suggested, a comparison of Weir's *Dead Poets Society* and Muriel Spark's novel, *The Prime of Miss Jean Brodie* (1961) – later filmed (Neame 1969), with the same title – would reveal the latter to be a 'counterpoint' to the former 'romantic pedagogy'. The musical metaphor, 'counterpoint', suggests that Spark's novel could be understood – or taught – in conjunction with Weir's film, a thought that is illuminating concerning both the differences and the similarities between the two narratives. Mr Keating's character, for instance, is the antithesis of a fascist (as the scene, where he uses the boys' tendency, to fall in step like marching soldiers for educational purposes, suggests) while Miss Brody is sympathetic to nationalism and fascism. Yet, both characters take 'education' of the young very seriously in the etymological sense of 'leading out', and both could be said to succeed with some of their students, while failing with others.

What Weir gets across via the character of Mr Keating is crucial for distinguishing fundamentally between human beings and AI, namely, that every human being is singular, unique, irreplaceable, and that this 'singularity' – *not* in Ray Kurzweil's (2006) sense of the term – consists (among other traits) in a person's capacity to follow their own, self-*chosen* path in life, where they are able to make a unique contribution to human culture and society. This is distinctively human. Human beings are not programmed in the sense that AI is; they have a certain genetic endowment (which some may conflate with being 'programmed'), but which parts of that endowment are activated depends on contingent things such as the specific environment (family, community, society, culture) in which a person grows up, including the specific people with whom they interact, such as loving, or, regrettably, sometimes abusive, parents, as well as 'good' or 'bad' teachers. The latter, in turn, have a ('positive' or 'negative') cultivating influence on a person, which shapes one's interests, and that either enhance – and are enhanced by – one's genetic predispositions, or do not enhance (and are not enhanced by) these. This convergence of genetic legacy and environmental factors have been thoroughly investigated and confirmed by researchers (see Cherry 2019; Francis and Kaufer 2011). The point is that every human being is a unique synthesis of – to use the customary terms – 'nature' and 'nurture', and on the basis of this synthesis a person makes his or her choices and decisions in life, becoming a (never-ending) 'work-in-progress'. Even if (once a person's personality is familiar to an observer) one's actions may

appear to be fairly predictable, this is never absolutely the case, given that everyone is able to deviate from what she or he has done in the past – as Albert Camus (1991: 10) famously pointed out, every person reaches a kind of 'borderline', beyond which they will not go. This, too, is distinctively human. In Camus's words (1991: 10):

> What is a rebel? A man who says no, but whose refusal does not imply a renunciation. He is also a man who says yes, from the moment he makes his first gesture of rebellion. A slave who has taken orders all his life suddenly decides that he cannot obey some new command …. Rebellion cannot exist without the feeling that, somewhere and somehow, one is right. It is in this way that the rebel slave says yes and no simultaneously. He affirms that there are limits and also that he suspects – and wishes to preserve – the existence of certain things on this side of the borderline. He demonstrates, with obstinacy, that there is something in him which 'is worth while ...' and which must be taken into consideration. In a certain way, he confronts an order of things which oppresses him with the insistence on a kind of right not to be oppressed beyond the limit that he can tolerate.

There are many historical as well as fictional instances of such rebels, who revolt in the face of something unbearably oppressive – the slave, Spartacus, who led a slave rebellion against the might of Rome until he was killed in battle in the year 71 BCE (BBC Ancient History: no date), and whose rebellion has been celebrated in various fictional works (see for example Fast 1951; and Kubrick 1960); Ché Guevara, who travelled through South America when he was a middle class medical student and, having witnessed the cruelty of capitalist mine owners towards their workers, rebelled by dedicating his life to fighting the injustices of capitalist exploittation, and many others (Biography: no date). While these examples illustrate the claim that the capacity to rebel or revolt is distinctively human, the question it raises is whether, in principle – if not in demonstrated fact – extant AI is capable of rebelling in the manner depicted in the science-fiction films referred to earlier. I have not been able to trace any such historical-factual AI-rebellion(s), but (as already intimated) am familiar with several fictional instances, which function as a kind of benchmark, or challenge, to AI-researchers: if it is indeed the case that AI-research is aimed at producing a

fully human simulacrum, then the challenge would be to construct an AI that is capable of rebelling. Among the science-fictional instances of such AI rebels there is the self-aware computer – that 'dinkum thinkum', called 'Mike' – in Robert Heinlein's classic tale of lunar rebellion, *The Moon is a Harsh Mistress* (1997), and then there are the Cylons in Moore's equally classic television series, *Battlestar Galactica*, that revolt against their human creators, just as the machines do in James Cameron's *Terminator* films, both paradigmatic instances that were referred to earlier. There are other examples of such fictional rebellions on the part of AI, but these will suffice to make the point. My reason for claiming that the capacity to rebel is a *challenge* to the AI-research community is parallel to what I have argued before (Olivier 2008) about the capacity to *love*, as well as the ability to make *moral decisions* – both of which are distinctively human, and are thematised in science fiction films, specifically Steven Spielberg's *AI – Artificial Intelligence* (2001) and Alex Proyas's *I, Robot* (2004), respectively. If an AI can be constructed that would demonstrate – as the AI (robotic) protagonists in both these films do (the robot-boy, David, in *AI*, and the robot, Sonny, in *I, Robot*) – its capacity to love, and to exercise moral choice (respectively), then AI will truly have succeeded in simulating a human being – instead of merely trying to copy, and surpass, human 'intelligence' in the narrow sense of rational, logical functioning, instead of considering human (mental and other) capacities in their entirety (more on this below).

Julia Kristeva gives one another perspective on 'revolt' by linking it with what is known as the 'pleasure principle' in psychoanalysis, in this way drawing attention to something else that is arguably distinctively human (Kristeva 2000: 7):

> Happiness exists only at the price of a revolt. None of us has pleasure without confronting an obstacle, prohibition, authority, or law that allows us to realize ourselves as autonomous and free. The revolt revealed to accompany the private experience of happiness is an integral part of the pleasure principle. Furthermore, on the social level, the normalizing order is far from perfect and fails to support the excluded: jobless youth, the poor in the projects, the homeless, the unemployed, and foreigners, among many others. When the excluded have no culture of revolt and must content themselves with ideologies, with shows and entertainments that far from satisfy the

demand for pleasure, they become rioters.

It is probably the case that, with the exception of the very few people who are familiar with psychoanalytic theory, including Kristeva's work, most people would be somewhat nonplussed by this excerpt. This is because she is not alluding to 'pleasure' in the usual sense of consumerist 'satisfaction' at having bought some or other 'high-end' consumer product such as a luxury German sedan, or the latest cell phone from Apple. To be sure, a certain variety of 'pleasure' accompanies the act of buying. This notwithstanding, it is not 'pleasure' in this everyday sense that Kristeva has in mind when she links 'revolt' with the 'pleasure principle', which was first formulated by Sigmund Freud. The pleasure principle, according to Freud, governs all human actions in so far as they are – consciously and unconsciously – aimed at removing obstacles to a state of relative psychic homeostasis or equilibrium (parallel to its biological or physiological counterpart). Such a state is never permanently achieved (which is why the expression 'relative homeostasis' is employed here) because that would be tantamount to death. In Kaja Silverman's words (1984: 56-57):

> According to Freud, the perception of deficiency triggers mental activity because of the psychic dominance of the 'unpleasure principle' (later termed the 'pleasure principle'). What this means is that the impulse to avoid unpleasure – i.e. to avoid an increase in tension – governs all psychic activity. When confronted with experiences that inspire excitation (i.e. tension), the mind attempts to substitute for them experiences that diminish that excitation. The consequent mental exertion is prompted by the wish for pleasure, and the condition of quiescence that follows in the event of success corresponds precisely to Freud's notion of pleasure. For Freud, pleasure represents the absence of unpleasure; it is a state of relaxation much more intimately connected with death than with life.

However, while Freud conceives of pleasure 'as a zero degree of tension' (Silverman 1984: 57) Kristeva associates it with 'revolt' because 'pleasure' in this sense can also be said to be fundamentally a function of resisting unjust or oppressive behaviour. Even if the latter is not immediately successful, the experience of justified resistance affords one 'pleasure' in so

far as one has not simply endured, or suffered, an untenable prohibition or groundless claim of authority over oneself – one has rebelled or revolted against it. In other words, what Kristeva considers to be the pleasure concomitant with 'revolt' is precisely the experience of having effectively challenged an obstacle of some sort, with good reason, such as some unjust law (like those reinforcing apartheid in the era that bore that name, or those, in the current era, which unfairly discriminate against minorities), or oppressive authority, in the name of something more valuable, or more imperative. In this respect it is notable that Kristeva uses the term 'revolt' in its etymological sense of 'returning' (2002: 100): 'The word revolt comes from a Sanskrit root that means to discover, open, but also to turn, to return'. Therefore, to rebel or revolt is to (re-)discover, in oneself, the psychic and moral resources to oppose, and possibly overcome the forces intent on dominating one. This capacity, I would argue, is another distinctively human attribute; one that an AI cannot possess in so far as it requires a personal history of linguistically oriented experience: one has to know what it means to be oppressed, or wronged, or unjustifiably persecuted, and to articulate this linguistically. However, AI is not only *factually* (that is, until now in its development) incapable of 'revolt' in the sense outlined above, but *in principle* incapacitated in this respect. I referred to Robert Heinlein's novel, *The Moon is a Harsh Mistress* (1997), and to the television series, *Battlestar Galactica*, as well as James Cameron's *Terminator* films above. All of these are science-*fictional* accounts of AI rebelling against the humans who created them, and just as the robots in Spielberg's *AI – Artificial Intelligence* and Proyas's *I, Robot* (mentioned earlier), show themselves as being capable of loving and exercising moral choice, respectively, in this way constituting a challenge to be actualised by AI-research, so, too, the imaginary AI that rebel against their human masters equally represent such a challenge. Arguably, it is a challenge that AI-research cannot meet, for reasons outlined above. Another film that instantiates a challenge of this magnitude is Alex Garland's *Ex Machina* (2014), in which the AI, or robot ('Ava'), not only succeeds in rebelling against 'her' human creators by outwitting them and deceiving them, but actually disposes of them before escaping from the residence where 'she' was kept. This film will be explored in greater detail below, but first a detour via a computer scientist's comparative evaluation of AI and human 'intelligence' is necessary to flesh out my claim, above, that there is something distinctive about being human.

A Computer Scientist on the (Human) Mind

In discussions of Artificial Intelligence (AI), one cannot but notice a general tendency, to restrict one's assessment of AI to its computing, calculating power, instead of considering a broader range of performative possibilities – something that was acknowledged by Sherry Turkle in her early work, *Computers and the Human Spirit*, although she herself was more interested in a computer's function as an 'evocative object, an object that fascinates, disturbs equanimity, and precipitates thought' (1984: 19). It is therefore worthwhile showing that it is not always the case that a comparison of humans with AI is carried out with reference to such a narrow conception of 'intelligence'. I shall do so by briefly referring to Yale University computer scientist, David Gelernter's challenge to his AI-research colleagues, whom he ac-cuses of (reductionist) 'computationalism', to pay attention to the full 'spectrum' of the human mind, instead. In his book, *The Tides of Mind: Uncovering the Spectrum of Consciousness* (2016; see also Olivier 2017), Gelernter uncovers this prevailing view of AI, predominantly in terms of its calculative capacity. Gelernter avers that most contemporary computer scientists, as well as so-called 'philosophers of mind' appear to believe that 'mind' (which coincides with their conception of AI) is confined to the logical, abstract, high-focus functions of what is thought of as 'rational' thinking (2016: xii-xiv). He raises the question, 'Why should philosophy of mind be obsessed with digital computers?', and provides the following answer (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.

Subsequent to this, however, it becomes evident that Gelernter believes it *not* to be the case. This stands in conspicuous contrast with Georg Schwarz's conviction, articulated in 1990, when the latter observed (rather optimistically) that (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* occurred with the publication of Gelernter's book – and by implication with that of many books, such as poetry volumes, where feeling and emotion form part of understanding (cognising) the poetry, before this book – insofar as it is precisely the 'non-computable [cognitive] functions' of the human mind on which he focuses. On the one hand, computationalism aims at demonstrating the 'parallel' functioning of the mind and the brain, all the while assuming that 'computing' – that is, calculation on the basis of a set of 'algorithmic' rules – is crucial to succeeding in this. On the other hand, Gelernter focuses on the *non-computable* cognitive functions of the mind, by identifying all those mental operations that computers, or AI and AI-research in their present guise (and arguably in *any* guise), cannot include under computation and computationalism. At present AI is known as 'weak AI' – that is, AI that depends on human programming for performing a variety of computational tasks. By contrast, AI researchers associated with the programme known as 'strong AI' are working towards AI that would supposedly be the equivalent of humans as conscious, thinking beings, and that would be capable of 'everything' that, and more than, humans can do (Armstrong 2017). One should note that, alongside the strong AI research programme, there is the escalating field of 'connectionism', which Medler (1998: 63) described, more than 20 years ago, as a theory of 'information processing' within the field of cognitive science – one that deviates from those hierarchical systems that utilise algorithmic rules for the manipulation of symbols. Instead, the cognitive models of connectionism concentrate on the development of 'parallel processing' similar to what takes place in the human brain, in this way mimicking its neurophysiology. However, whatever the differences between these two research programmes might be, in my view they share the reductionist principle of equating the mind with a kind of software and the brain with hardware, which is untenable – particularly, but not only (see for example the phenomenological work of Maurice Merleau-Ponty and Jean-Francois Lyotard on embodiment; Olivier 2002b) in light of Gelernter's work. My reasons for this claim are as follows.

It is unsurprising that philosophers of mind have noticed the resemblance between the functioning of the human brain and that of computers (Gelernter 2016: xviii-xix), where the mind is compared to software and the brain to hardware. After all, human beings designed the computers to perform *some* of the functions their minds usually perform, but *not all of these functions*. Importantly, this does not logically imply that minds *are* computers; for instance, *disembodied* computers do not have 'life-stories', while humans, as '*embodied* minds', do; the fact of embodiment is all-important here, given that a person's sense of their own life-history unavoidably includes sensory memories, such as an aroma, or a taste, that are dependent upon having a (and in a certain sense 'being' one's) body. Add to this that digital computers operate (Gelernter 2016: xiii-xiv) by way of '*processes*' ('actions-in-time'), in the guise of 'following a prearranged set of rules' or a series of prescribed steps, from which they do not deviate, and another contrast between them and humans emerges. This pertains to an algorithmic function of AI. Considering the theme of this paper, it is significant that such rule-based functioning, on the part of AI, is qualitatively different from humans' ability – if not the unavoidable intermittent occurrence – of deviating from 'prescribed rules', whether these are the rules of a sport, like soccer or tennis, or the rules for writing examinations. People are not algorithmically programmed beings. In the light of this manifestation of either fallibility (when deviating from a rule is not done deliberately), or wilfulness (when rules are deliberately ignored), anyone who attributes to humans, analogous to AI, a primary 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. To be sure, some thinkers defend the latter position, but – as Kant (1960: 30) demonstrated in the late 18th century – unless one presupposes freedom of the will on the part of human beings, talking about ethical or moral choice is nonsensical because, without freedom of the will, all apparent 'morally good' (or, for that matter, 'immoral') human actions would be predetermined by some heterogeneous 'law' or power.

The assertion on the part of philosophers of mind that human brains (which putatively use 'software' called 'minds') *are* indeed computers, albeit 'encased' differently (in fact, 'embodied', which no computer hardware or software is), appears to be what really interests Gelernter. This position is what gave rise to 'computationalism' (2016: xviii), which rests on the

following argument (more or less): Because computing is a form of thinking, namely, clear, rational thinking, and computers compute, computers can be called 'thinking machines'. And studying the theory, performance and structure of digital computers amounts to studying the mind, because thinking is something done by minds. It seemed to make sense, therefore, that the 'intelligent' performance of 'rational', thinking tasks can be carried out by computers or AI. Hence the philosophical field of 'computationalism'.

Contrary to the majority of other computer scientists, in this carefully argued book Gelernter establishes a case for the irreducible difference between 'brain' and 'mind'. He does this, among other ways, by elaborating on the distinctive kind of mental activity known as 'free association', in contrast to conscious, focused mental activity, as well as on the vital contribution of fantasy and dreaming to creative thinking. In sum, he examines all those mental activities in which human beings habitually engage in the course of everyday living. One is struck by the irony that a philosophically minded computer scientist such as David Gelernter tackles the daunting task of demonstrating 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', as the title of the book (2016) suggests. As a computer scientist his voice is refreshingly different in a world where there is a growing proclivity, particularly among computer scientists and philosophers of mind, to use something conceived and built by human beings, namely the computer, *reductively* and retrospectively as a model to comprehend what it is to be human.

Gelernter reminds his readers forcibly about the full scope ('tides') of the human mind, which is nowhere reflected in current AI research or development. To uncover these 'tides', Gelernter consults the works of literary geniuses like Shakespeare, Proust, Tolstoy and South African novelist, J.M. Coetzee. It is telling – given the fact that something other than the 'conscious' mind is also involved in creativity – that Gelernter gives significant attention to that figure who (more than anyone else, perhaps) changed the way human beings think about themselves – Sigmund Freud, the inventor of psychoanalysis as a novel discipline – to be able to gauge the mind's true 'depth'. After all, the fruits of mental accomplishments across a whole 'spectrum' of activities have to be scrutinised. It is clearly not sufficient to restrict one's investigation to the logical, focused functions of so-called 'rational' thinking, as most philosophers of mind and computer scien-

tists do. Instead, Gelernter considers the mind across what he calls a 'spectrum', which ranges from 'high focus' mental engagements, like intensely self-aware reflection, through 'medium' activities such as experience-related thinking (for example emotion-suffused daydreaming), to 'low focus' functions such as 'drifting' thought, with accompanying emotions burgeoning, as well as to dreaming (2016: 3; 241-246). Moreover, the different functions of memory are revealing about what it means to be human, as opposed to artificially intelligent. Gelernter argues that, at the 'high focus' level of the mental spectrum, memory is utilised in a disciplined way, compared to which, at the medium-focus niveau it 'ranges freely', while the low-focus level is characterised by memory that 'takes off on its own'. Needless to stress, 'memory' in this multivalent sense is something that the narrow sense of 'memory' on the part of AI does not encompass, and even if AI that is capable of drawing on memory at different levels could be developed, it would be insurmountably hampered by the fact that it lacks a body- and time-rooted life-history, of which multi-level human memory is the repository.

From what we have learned about this range of 'tides' characteristic of the human mind (most of which are ignored by computationalism) it should be clear that Gelernter diverges from what computer scientists who subscribe to futurologist Raymond Kurzweil's techno-optimism (2006: 39-46) believe. According to Kurzweil, humanity finds itself 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 AI will immeasurably surpass human intelligence, and pave the way for humans to merge with machines. More specifically, against the backdrop of the discussion of Gelernter's elaboration on the human mind, as opposed to the 'AI mind' – if there is such a thing – one can safely say that, for him, 'intelligence' in the narrow 'computationalist' sense is not the only thing to consider when it comes to AI emulating what it is to be human. Moreover, there are other human capacities, some of which will be further discussed below, which AI would have to acquire if it is to be a 'proper' simulation of a human being (see Olivier 2008 and 2017 for an elaboration on other such differentiating human attributes). What one learns from this intellectually circumspect computer scientist (Gelernter) is that the human mind is multi-facetted, instead of being the impoverished, unidimensional faculty that computationalism reduces it to. The different 'tides' which he discerns in and of the mind, *all* belong to it *irreducibly*, instead of only the one that is located

at the level of logical, 'high focus' mental activity. By concentrating exclusively on the latter, conventional AI-research has impoverished the appropriately encompassing, multi-dimensional conception of the mind as one knows it from the arts and even from daily experience. Following Gelernter, it is misguided (as computationalism encourages one) to believe that the high-focus level of mental functions *alone* is what 'mind' is, whether in its AI-guise or its human embodiment (2016: xi-xix).

Furthermore, as far as the nature of *creativity* is concerned – as it is displayed on the part of the towering creative thinkers (like Shakespeare) referred to in his book – such a one-dimensional conception of the mind cannot possibly do justice, in Gelernter's view, to this distinctive human ability. In the final analysis this marks an insuperable difference between AI, on the one hand, and creative human intelligence, on the other. The instances of AI 'composing' music that one has witnessed of late do not negate this argument. Such supposedly 'creative' musical composition is unthinkable without pre-programming the AI in question to arrange musical sounds on the basis of certain algorithms, as a brief perusal of currently available AI composition software shows (Wondershare Filmora 2020) – despite the promise of 'original' musical compositions, it is clear that the effective use of such software depends on human creativity. For example, under 'Amper Music' it is stated that:

> Amper Music is a cloud-based platform designed to simplify the process of creating soundtracks for movies and video games, as it produces AI generated algorithms that help users create music in a variety of music genres.

In addition to the distinctive human attributes which have been discussed so far, one cannot overlook moral or ethical capacity and agency. In this regard, the following observation by Gelernter is highly pertinent to the question of AI and ethical consciousness, which will be addressed below. This is all the more significant if it is kept in mind that computalionalism equates AI and (the human) mind (Gelernter 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.

Put differently, the province of logic coincides with that of AI as conceived by computationalism (but only *partly* with the human mind); it is therefore devoid of moral significance. On the other hand, narratives – belonging to a realm completely alien to that of impersonal logic – are usually shot through with moral significance. This has far-reaching implications as far as challenges to the AI-research community are concerned. AI (constructed according to the computationalist model of the mind) is virtually unsurpassable at the level of logic, *but crucially*, ethically meaningful action is fundamentally incommensurate with its capabilities. In the light of this, I am always surprised that writers who point out the irreconcilable differences between human beings and AI generally seem to concentrate on the *cognitive* functions of the mind in a narrow sense – this is even largely true of someone like Gelernter, who shows great sensitivity for the entire 'spectrum' of the mind. Yet, this happens largely to the exclusion of human ethical and moral traits that are outside of the functioning of AI in computationalist terms, and arguably even in principle.

It is therefore strange that, although he makes the distinction between logic and a morally meaningful story, Gelernter does not focus explicitly and in a sustained manner on this irreducible difference between human 'intelligence' in the most inclusive sense of the word, and AI. In this encompassing sense, human 'intelligence' arguably includes the abilities of making ethical (and also aesthetic) decisions, let alone the capacity of loving someone (Olivier 2008). This is the case because a moral judgment of a person being a thief, or a murderer, is not in itself a 'cognitive judgement' of the scientific variety, susceptible to scientific theory-guided interpretation. And yet, it does presuppose 'knowledge' of a certain kind, namely ethical, moral, or 'practical' knowledge (known as 'praxis'). 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, and implied by his distinction between scientific logic and religious narrative. In what follows I would like to demonstrate the significance of Gelernter's discernment of moral significance in stories by focusing on a paradigmatic cinematic narrative which, simultaneously, represents a challenge that, in my humble judgement, AI-research cannot meet.

Garland's *Ex Machina* (2014), Anthro-pessimism and Deceptive AI

To test the claims of AI-research and development – that an AI-equivalent (or simulation) of human 'intelligence' can be produced – consider Alex Garland's 2014 science fiction thriller, *Ex Machina* (tellingly subtitled 'What happens to me if I fail your test?' and in some versions: 'There is nothing more human than the will to survive'), where what one might call an anthro-pessimistic view of humans is given a science-fictional twist, ascribing a strangely 'human' mode of thinking and acting to the 'female' AI character, or 'fembot'. In fact, given what it, or 'she' is willing to do to ensure her own survival, one might say that her creator, Nathan Bateman (Oscar Isaac) has succeeded only too well in recreating human 'nature' in the robot, named 'Ava' (Alicia Vikander), with its echo of the mythically 'first' woman, Eve. The lesson from this, as I shall argue below, is that artificial intelligence researchers should perhaps not aim too squarely for a human simulacrum in their work, lest they succeed, because *Ex Machina* shows one that, if they do, one has reason to be very pessimistic about the future relations between people and their robotic creations. (This is reminiscent, once again, of the mammoth television series, *Battlestar Galactica*, where the anthropomorphic robots, called Cylons, turn on their human masters with the intent to destroy them, and very nearly succeed (see for instance Olivier 2015).

The film's narrative (which must be briefly reconstructed to be able to make my point about the challenge it poses to AI research) concerns Nathan's wish, to demonstrate beyond any doubt that he has achieved the ultimate goal of AI-development, namely, to create a being (Ava) that is a human simulacrum in all respects, including the capacity to evoke the personal, emotional and even romantic interest of a human being. With this in mind, Nathan invites a promising employee, Caleb Smith (Domhnall Gleeson), a programmer at his software company, Blue Book, to his secluded luxury retreat. Here he instructs Caleb to meet with Ava regularly to judge his own ability of entering into a 'relationship' of sorts with 'her', regardless of her artificial status. Not surprisingly, Caleb enjoys his meetings with Ava, who acts and converses like an intelligent human being, and despite his awareness that she is a robot he grows to like and trust her, which should hardly be difficult, given her beautiful feminine face and shape, although

transparent parts of her body reveal its artificial android structure. Confirming her human simulacrum-status in emotional terms (already posing a challenge to AI research), Ava goes as far as confessing to Caleb that she is strongly attracted to him. Moreover, Caleb learns from her that she wishes to escape from her confined existence to the outside world, and having witnessed increasing manifestations of disturbing narcissistic behaviour on Nathan's part, he becomes amenable to her desire.

However, because Caleb has become aware of Nathan covertly eavesdropping on their meetings through technical surveillance, they cannot converse openly, until Ava reveals that she can cause power failures that interrupt his surveillance system, as well as initiating the locking of all doors by the security system. During one of these meetings Ava urges Caleb not to trust Nathan. At this point, Caleb has already seen the models that preceded Ava and failed Nathan's stringent 'Turing Tests', which led to their 'termination'. To Caleb's alarm, Nathan shares with him his intention to repro-gramme Ava, which, in effect, would terminate the Ava he knows – what replaces her would be a new AI. With the intention of rescuing Ava, Caleb tricks Nathan into too much drinking and subsequently passing out. He gains access to the latter's computer with his security card and modifies the security system by changing relevant code(s). Having grown suspicious that he may himself be a robot – after witnessing revealing (and alarming) video material of Nathan's actions towards decommissioned robots – Caleb attempts to verify his own flesh-and-blood status by cutting himself. Together with Ava he hatches a plot to repeat his neutralising of Nathan and re-coding of the system to open the doors instead of closing them during a power failure, so that they can escape together. However, their plans are thwarted by Nathan, who has listened to all their putatively secret conversations by means of an independently powered video-camera. In the end it transpires that the humanoid robot's plan has been all along to escape at all costs, even if this means 'cynically' eliminating the human being who has assisted 'her'. Her ruthless escape entails, firstly, killing Nathan (with the help of another android, Kyoko – whose job it was to see to the satisfaction of Nathan's sexual needs – but who is destroyed in the process), and covering the trans-parent parts of her body with 'skin'-parts from earlier robots, to resemble a human being perfectly. But the finishing touch of her cynicism and disingenuousness comes when she abandons Caleb, who was rendered unconscious by Nathan, and wakes up just in time to see her

escaping to the outside, while he remains trapped inside the securely locked house, probably with no means of getting out. The film concludes with Ava completing her escape by boarding the helicopter intended for Caleb, presumably heading for human society, judging by the closing image-sequences in the film.

The most surprising development in the plot is undoubtedly when Nathan informs Caleb that, in his view, Ava has been manipulating him with the aim of using him to escape – her 'romantic' interest in him having been a mere pretence, according to Nathan. But (and this is the significant element for the present paper) precisely *this* has been the true test of her success as a human simulacrum – the fact that she has the intelligence as well as *the moral unscrupulousness to manipulate a human being for her own hidden purposes*. Needless to stress, this is an extremely pessimistic, if not downright cynical, assessment of what makes us human beings. To illustrate, in terms of Jürgen Habermas's theory of discourse ethics, instead of having practised 'communicative action' (where one is as open and sincere as possible in your communication with others) in her conversations with Caleb, Ava has been indulging in 'strategic action' instead (where one, while *ostensibly* engaging in 'communicative action', has a hidden agenda, hiding one's true intent for the sake of manipulating others). Lasse Thomassen comments as follows on Habermas's crucial distinction (2010: 68).

> It is important to highlight the essential difference between communicative and instrumental/strategic action. Where the latter are oriented towards success in the non-social and social worlds respectively, the former is oriented towards reaching understanding in the social world ….

Clearly, in the light of this distinction Ava was not interested in 'reaching understanding' with Caleb, despite giving the impression that this was indeed the case; her communication with him was 'oriented towards success in the non-social and social worlds', where 'success' denotes her ingeniously engineered escape from the confines of a technologically experimental space. In other words, she simultaneously succeeded specta-cularly at what must be regarded – as confirmed by Habermas – as something distinctively human: the ability to act 'strategically', which is here a euphemism for 'deceptively'.

Conclusion: 'Intelligence' and Duplicity

This is the central message of Garland's film, then: artificial intelligence researchers will have succeeded in constructing an android – a robot that perfectly simulates human behaviour – when they can come up with something (or 'someone') as duplicitous or devious as Ava, who does not hesitate to use others (in this case Caleb) to reach their own dubious goals. One might argue – correctly, I believe – that Ava was merely doing what was necessary to survive, instead of being 'terminated' by Nathan, her 'creator'. The important point, however, is that in doing so she passed the rigorous 'Turing Test' that he had devised for her in relation to Caleb in the role of assessor with flying colours: she had become indistinguishable from lying, deceitful humans. After all, it would hardly be necessary for a thinker like Habermas – or any thinker before him who contributed to the discipline of ethics – to distinguish between two kinds of discursive 'action' (the one ethically acceptable; the other ethically dubious) unless humans routinely showed themselves capable of both. Succinctly phrased: there would be no need for ethics unless humans acted unethically (or were capable of doing so). Garland's film demonstrates, in science-fictional terms, that such (duplicitous, and therefore unethical) action would be a failsafe test for the development of AI that could fool anyone into believing it was 'properly' human, and not only in appearance. Finally, even David Gelernter could learn from the insights embodied in this cinematic work; that it is not sufficient to enumerate all the various 'tides' on the spectrum of the human mind to show where computationalist AI research falls short; other human capabilities, such as the ability to deceive, are equally valid as distinguishing attributes when compared to AI.

References

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

BBC Ancient History n.d. Spartacus: http://www.bbc.co.uk/history/historic_figures/spartacus.shtml (Accessed 6 October 2020.)

Biography n.d.): Ché Guevara: https://www.biography.com/political-figure/che-guevara (Accessed 6 October 2020.)
Cameron, J. (Dir.). 1984. *The Terminator*. United States: Orion Films.
Cameron, J. (Dir.). 1991. *Terminator II – Judgement Day.* United States: Columbia Tristar.
Camus, A. 1991. *The Rebel. An Essay on Man in Revolt.* Bower, A. (trans.). New York: Vintage.
Cherry K. 2019. The Age-old Debate of Nature vs. Nurture. Medically Reviewed by a Board-certified Physician. Available at: https://www.verywellmind.com/what-is-nature-versus-nurture-2795392 (Accessed on 20 July 2019.)
Fast, H. 1951. *Spartacus.* New York: Blue Heron Press.
Garland, A. (Dir.) 2014. *Ex Machina.* United Kingdom: Film4. https://doi.org/10.5040/9780571343041-div-00000006
Gelernter, D. 2016. *The Tides of Mind: Uncovering the Spectrum of Consciousness.* New York: Liveright Publishing Corporation.
Heinlein, R.A. 1997. *The Moon is a Harsh Mistress.* New York: Orb Books.
Francis, D. & D. Kaufer 2011. Beyond Nature vs. Nurture. *The Scientist.* October issue. Available at: https://www.the-scientist.com/reading-frames/beyond-nature-vs-nurture-41858 (Accessed on 3 April 2019.)
Kant, I. 1960. *Religion within the Limits of Reason Alone.* Greene, T.M. and H.H. Hudson (trans.). New York: Harper Torchbooks.
Kristeva, J. 2000. *The Sense and Non-sense of Revolt: The Powers and Limits of Psychoanalysis.* Volume I. Herman, J. (trans.). New York: Columbia University Press.
Kristeva, J. 2002. *Revolt, She Said: An Interview by Philippe Petit.* (Rainer Ganahl & Rubén Gallo.) O'Keeffe, B. (trans.). New York: Semiotext(e).
Kubrick, S. (Dir.) 1960. *Spartacus.* USA: Universal International.
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 July 2016.)
Medler, D.A. 1998. A Brief History of Connectionism. *Neural Computing Surveys* 1: 61-101. http://www.blutner.de/NeuralNets/Texts/Medler.pdf (Accessed 23 May 2017.)
Moore, R.D. (Dev.) 2004 - 2009. *Battlestar Galactica.* United States: NBC Universal Television.

Neame, R. (Dir.) 1969. *The Prime of Miss Jean Brodie*. United Kingdom: 20th Century Fox.

Olivier, B. 2002. Time, Technology, Cinematic Art and Critique in *The Terminator* and *Terminator II – Judgment Day*: A Philosophical Interpretation. In *Projections: Philosophical Themes on Film.* Second, enlarged edition. Port Elizabeth: University of Port Elizabeth Publications.

Olivier, B. 2002a. Reason and/or Imagination? Peter Weir's *Dead Poets Society*. *Film and Philosophy* 5/6, (US *Journal of the Society for the Philosophic Study of the Contemporary Visual Arts*) February; General interest edition: 14 - 24. https://doi.org/10.5840/filmphil20025/63

Olivier, B. 2002b. 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. 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, September: 30 - 44. Available at: http://www.film-philosophy.com/index.php/f-p/article/view/56/41; https://doi.org/10.3366/film.2008.0014

Olivier, B. 2015. *Battlestar Galactica*, Technological Development and Eternal Recurrence. *Film and Philosophy* (US *Journal of the Society for the Philosophic Study of the Contemporary Visual Arts*) 19: 127 - 140. https://doi.org/10.5840/filmphil2015199

Olivier, B. 2017. Artificial Intelligence (AI) and Being Human: What is the Difference? *Acta Academica* 49 (1), pp. 2-21. https://doi.org/10.18820/24150479/aa49i1.1

Proyas, A. (Dir.). 2004. *I, Robot.* United States: 20th Century Fox.

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

Scott, R. (Dir.). 1982. *Blade Runner.* United States: Warner Bros.

Silverman, K. 1984. *The Subject of Semiotics.* Oxford: Oxford University Press.

Spark, M. 1961. *The Prime of Miss Jean Brodie*. London: Macmillan.

Spielberg, S. (Dir.). 2001. *AI – Artificial Intelligence.* United States: Warner Bros.

Thomassen, L. 2010. *Habermas: A Guide for the Perplexed.* New York: Continuum.

Turkle, S. 1984. *Computers and the Human Spirit.* Cambridge, Mass.: The MIT Press.

Weir, P. (Dir.). 1989. *Dead Poets Society.* United States: Touchstone Pictures.

Wondershare Filmora 2020. *Top 10 AI Music Composers in 2020.* https://filmora.wondershare.com/audio-editing/best-ai-music-composer.html (Accessed 29 October 2020.)

Bert Olivier
Department of Philosophy
University of the Free State
South Africa
OlivierG1@ufs.ac.za
bertzaza@yahoo.co.uk