More Human Than All Too Human: Challenges in Machine Ethics for Humanity Becoming a Spacefaring Civilization

Guy Du Plessis[1]

1 Utah State University

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

It is indubitable that machines with artificial intelligence (AI) will be an essential component in humans' quest to become a spacefaring civilization. Most would agree that long-distance space travel and the colonization of Mars will not be possible without adequately developed AI. Machines with AI have a normative function, but some argue that it can also be evaluated from the perspective of ethical norms. This essay is based on the assumption that machine ethics is an essential philosophical perspective in realizing the aim of humanity becoming a spacefaring civilization. In this essay, I explore two questions in the field of machine ethics, that I believe to be relevant to the role AI will play in long-distance space travel. The first is, should moral theory be extended to include machines with AI, and second, can machines be fully ethical agents? In this essay, I define AI and then discuss the difference between implicit, explicit and full ethical agents in relation to machines with AI. I then present the argument that the inclusion of moral theory is essential in the development of machines with AI. Without an adequate inclusion of moral theory in the design of AI it may pose an existential threat to humanity, especially in the development of super-intelligent machines. I also highlight that conceptual clarity is essential in the field of machine ethics and the choice of the conceptual foundation that informs AI research and development has ethical implications, especially in the case of super-intelligent machines. This essay is an exploratory and speculative philosophical analysis of certain aspects of machine ethics relevant to long-distance space travel and does not attempt to provide definitive answers to the questions posed in the essay, but instead aims to bring attention to what I deem important considerations.

> *"More human than human, is our motto."*
> *(The Tyrell Corporation, Blade Runner)*
>
> *"Everything is only–human, all too human?"*
> *(Friedrich Nietzsche, Human, All Too Human: A Book for Free Spirits)*

Introduction

Machines with artificial intelligence (AI) are ubiquitous and found in medical equipment, satellites, internet search engines, the internet of things (physical objects embedded with sensors and other technology to connect with other devices and systems via the internet and exchange data), mobile phones, video games, etc. AI applications are vital for security, surveillance, and medical science. Stuart Russell and Peter Norvig (2009) define AI as machines that imitate the mental functions of humans (problem-solving and learning), and which can "[think] humanly" (p. 2).

It is indubitable that machines with AI will be an essential component in humans' quest to become a spacefaring and multi-planetary civilization. Most would agree, including Elon Musk (2018), that long-distance space travel and the colonization of Mars will not be possible without adequately developed AI. Machines with AI have a normative function, but some argue that it can also be evaluated from the perspective of ethical norms (Moor, 2006). This essay is based on the assumption that machine ethics is an essential philosophical perspective in realizing the aim of humanity becoming a spacefaring civilization.

The notion of machine ethics is a controversial topic. Even its existence is debated. James Moor (2006, p. 2) highlights this by stating that

> *[t]he question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions.*

In this essay I explore questions around machine ethics, i.e., can machines be full ethical agents, and should moral theory be extended to include them?

I begin my essay by defining artificial intelligence (AI) and then discuss the difference between implicit, explicit and full ethical agents in relation to machines with AI. I present the argument that the inclusion of moral theory is essential in the development of machines with AI (a similar argument is made by James Moor in his article *The nature, importance, and difficulty of machine ethics*, 2006, and by David Gunkel in his article*A Vindication of the Rights of Machines,*2013). Without an adequate inclusion of moral theory in the design of AI it may pose an existential threat to humanity, especially in the development of super-intelligent machines in "a future period during which the pace of technological change will be so rapid, its impact so deep, that human life will be irreversibly transformed" (Kurzweil, 2015, p. 146). I also highlight that conceptual clarity is essential in the field of machine ethics and the choice of the conceptual foundation that informs AI research and development has ethical implications, especially in the case of super-intelligent machines that will be used in long-distance space travel. An example of the role of super-intelligent machines in long-distance space travel and its existential and ethical implications are exemplified in the film *2001: A Space Odyssey*. I will discuss this in greater depth later, but for now, it is sufficient to say that adequately designed AI will determine the success or failure of humanity attempting to become a spacefaring civilization – and the adequacy of AI design for the purpure of long-distance space

travel will greatly be influenced by our choice of conceptual and ethical foundations.

This essay is an exploratory and speculative philosophical analysis of certain aspects of machine ethics relevant to long-distance space travel and does not attempt to provide definitive answers to the questions posed in the essay, but instead aims to bring attention to what I deem important considerations.

Defining Artificial Intelligence

In this essay, I draw a distinction between weak and strong AI, as defined by John Searle (1980). According to Bishop (2021), weak AI "focuses on epistemic issues relating to engineering a simulation of human intelligent behavior" (p. 2). Weak AI is limited to specific functions and is found in voice-based virtual assistant technologies such as Alexa and Siri. Weak AI gives the appearance of intelligence without real understanding or true cognitive states.

In the 1956 Dartmouth workshop, strong AI research was succinctly formulated as a machine that can simulate every aspect of human intelligence, because we are able to so precisely describe human learning and intelligence (Müller & Bostrom, 2014). According to John Searle (1980), strong AI postulates that a computer or machine can be programmed to have understanding in the same way as humans and have similar cognitive states. In strong AI a machine would have self-awareness and consciousness and be able to learn, solve problems and plan for the future. Strong AI is associated with artificial general intelligence, human-level intelligence, superintelligence and artificial consciousness (Braidotti, 2013). Ultra-intelligence is another term used for "a machine that can far surpass all the intellectual activities of any man, however clever" (Good, 1965, p. 33). Strong AI aims to create intelligent machines indistinguishable from human intelligence. Currently, strong AI is only hypothetical, and far from realization. Alan Turing (2009) developed a test to assess whether a computer's behaviour could be differentiated from human behaviour. He explains that an "interrogator" tries to differentiate between human-generated output and computer-generated output via a series of questions. If the interrogator fails to distinguish the two, the computer is seen as having human-like intelligence and is according to Searl (1980) an example of a strong AI.

There are two theories that often inform the trajectory of strong AI research: the social construction of technology (SCOT) and technological determinism. Proponents of SCOT believe that human activity is not determined by technology, but that human action instead shapes technology. SCOT suggests that "the natural world has a small or non-existent role in the construction of scientific knowledge" (Collins, 1981, p. 3). Moreover, SCOT theorists argue that understanding how technology is used must include comprehension of its social milieu. The SCOT model is an example where theoretical assumptions inform models of mind, in turn informing AI, and if misguided inevitably leading to failure. I agree with Maze (2001) on the self-contradictory nature of (strong) social constructionism: "[S]ocial constructionist metatheory allows that any coherent epistemology must be self-reflexive, but while it denies that any assertion can be true, and that there are any independent realities to be referred to, it nevertheless treats discourse as having objective existence, and assumes that its own statements about discourse are true. Thus, in asserting its own basic premise it contradicts it" (p. 393).

In contrast, technological determinism is based on the premise that cultural values and social structures are determined by

technology. Determinism provides a reductionist view of how societies develop. It could be argued that the most well-known adherent of a technological determinist position was German philosopher Karl Marx. He argued that technologies, and specifically the means of technological production, are the primary determining factor influencing human relations and organizational structures: the technological and economic base is the primary factor determining societal development and social relations. One ontic aspect – technology and economics – is prioritized. Marx believed he could predict a society's development by analyzing its means of production and economic base. Marx is perhaps better viewed as a historicist – a term that is undeserving of praise, according to Popper et al. (2002).

In conclusion, Haugeland (1985) provides a useful definition, alluding to a foundational supposition that underpins strong AI research

> *The fundamental goal is not merely to mimic intelligence or produce some clever fake … AI wants only the genuine article:* ***machines with minds*** *… This is not science fiction, but real science, based on a theoretical conception as deep as it is daring: namely [that], we are, at root, computers ourselves. (p. 2)*

In the next part of the essay, I discuss machines as ethical-impact agents. Much of the discussion of machine ethics relates to strong AI, but it is also relevant to weak AI, especially in the case of AI which is considered an implicit ethical agent.

Ethical-Impact Agents

According to Michael Anderson & Susan Anderson (2007), a goal of machine ethics is to create machines that are guided by a set of ethical principles when making decisions and performing their normative functions. According to James Moor (2006), machines can be ethical-impact agents, and he makes a distinction between an "implicit ethical agent" and an "explicit ethical agent."

Machines that are implicit ethical agents are ones that have been programmed by a human to follow ethical behaviour or avoid unethical behaviour, but they cannot represent ethics explicitly. Its behaviour and decision-making are largely deterministic and restricted by its programming. Therefore, the machine does not decide what is ethical behaviour, this is predetermined by the human programmer.

On the other hand, machines that are explicit ethical agents can use ethical principles in calculating the best course of action when confronted with an ethical dilemma. Simply put, machines that are explicit ethical agents can represent ethics explicitly. According to Moor (2006), most of the interest in the field of machine ethics is aimed at creating machines that are explicit ethical agents.

According to Anderson & Anderson, (2007) what is critical in the implicit vs explicit ethical agent distinction, is not merely who makes the ethical decision, the machine or the programmer, but rather the machines' capacity to justify ethical

judgements “that only an explicit representation of ethical principles allows” (p. 2). An explicit ethical agent can appeal to an ethical principle and explain why an action is right or wrong. In contrast, an implicit ethical agent does not apply ethical principles in its decision-making and can be seen as lacking something essential for it to be considered an ethical agent. Immanuel Kant (1785) made a similar distinction when highlighting the difference between an agent that acts according to duty and one that consciously follows an ethics principle.

Can Machines be Full Ethical Agents?

A typical human that is not under the influence of psychoactive substances is normally considered a fully ethical agent. For a human or machine to be considered a full ethical agent, it needs to be able to make explicit ethical decisions and must have the capacity to justify them. According to Moor (2006), for humans to be full ethical agents they need to have consciousness, intentionality, and free will. In his article *The Nature, Importance and Difficulty of Machine Ethics*, Moor (2006) presents the question of whether a machine can be considered a fully ethical agent. This is a question that has typically led to heated debate and conflicting points of view. Moor (2006) highlights that for many there is a clear distinction between the ontological nature of machines and humans, and this clear distinction prevents machines from being considered to be full ethical agents. Moor argues for even if machines are not full ethical agents, they can still be considered to have an ethical impact. I agree with Moor and a graded ethical agent continuum for machines may be a better scale to use than a binary yes or no approach.

Whether machines can be considered full ethical agents is a difficult question to answer, and until we develop the technology that allows us to create machines that have the capacity to become full ethical agents a definitive answer may continue to elude us. Yet, this should not prevent us from exploring this question and examining its philosophical and real-world consequences. Like others I would argue that we should have some type of safeguard that would prevent machines from behaving unethically, whether we consider them implicit, explicate or full ethical agents. Especially, in the case of the development of super-intelligent machines, ethical safeguards are imperative.

There are several reasons why the question of whether machines can be considered full ethical agents is difficult to answer. One reason is conceptual. There is no conceptual clarity regarding who can be considered an ethical agent as well as for the phenomenon (consciousness, intentionality, free will) that Moor points out is needed for a human or machine to be considered a full ethical agent.

All theories have ontological and epistemological ancestry or foundational assumptions, implicit or explicit (Slife, 2005). Consequently, conceptions of mind that guide AI research and development, are based on philosophical assumptions and metatheoretical foundations, which influence the trajectory of development and research. Hence attempts to mechanize human-level reasoning and functioning will not be actualized if guided by a deficient conceptual foundation, and consequently, the project of strong AI is then doomed to fail, or if an artificial mind or super-intelligent computer is developed, it might result in undesirable activity, with disastrous consequences.

In the history of moral philosophy questions related to moral standing typically focused on agency and the figure of the

moral agent. “Virtue ethics, and Greek philosophy more generally,” Luciano Floridi explains (1999, 41), “concentrates its attention on the moral nature and development of the individual agent who performs the action. It can therefore be properly described as an agent-oriented, ‘subjective ethics.’” Gunkel (2014, p. 3) argues that

> *[w]hen considered from the perspective of the agent, ethics inevitably and unavoidably makes exclusive decisions about who is to be included in the community of moral subjects and what can be excluded from consideration…But who counts—who, in effect, gets to be situated under the term “who”—has never been entirely settled, and the historical development of moral philosophy can be interpreted as a progressive unfolding, where what had once been excluded (i.e., women, slaves, people of color, etc.) have slowly and not without considerable struggle and resistance been granted access to the gated community of moral agents and have thereby also come to be someone who counts. Despite this progress, which is, depending on how one looks at it, either remarkable or insufferably protracted, there remain additional exclusions, most notably nonhuman animals and machines.*

Machines in particular have been understood to be mere artifacts that are designed, produced, and employed by human agents for human-specified ends. This instrumentalist and anthropocentric understanding has achieved a remarkable level of acceptance and standardization, as is evident by the fact that it has remained in place and largely unchallenged from ancient to postmodern times.

> *Beginning with the animal rights movement, however, there has been considerable pressure to reconsider the ontological assumptions and moral consequences of this legacy of human exceptionalism. Extending consideration to these other previously marginalized subjects has required a significant reworking of the concept of moral agency. For this reason, the question of moral agency has come to be disengaged from identification with the human being and is instead often referred to and made dependent upon the generic concept of “personhood” Gunkel (2014, p. 3).*

Scott (1990, p. 7) states that “[t]here appears to be more unanimity as regards the claim that in order for an individual to be a moral agent s/he must possess the relevant features of a person; or, in other words, that being a person is a necessary, if not sufficient, condition for being a moral agent.” The problem is that there is little or no agreement on what constitutes a person, and the literature on the subject is cluttered with varied formulations and frequently conflicting criteria. As Daniel Dennett (1998, p. 267) writes, “One might well hope, that such an important concept, applied and denied so confidently, would have clearly formulatable necessary and sufficient conditions for ascription, but if it does, we have not yet discovered them. In the end there may be none to discover. In the end we may come to realize that the concept of person is incoherent and obsolete.”

In order to deal with this conceptual problem, researchers frequently concentrate on the one "person making" quality that appears on most lists of "personal properties," whether they include just a couple of simple elements (Singer 1999, p. 87) or involve numerous "interactive capacities" (Smith 2010, p. 74) — and one such property is consciousness. “Without consciousness there is no person,” said John Locke (1996, p. 146). Or as Kenneth Himma (2009, p. 19) argues, “moral

agency presupposes consciousness…and that the very concept of agency presupposes that agents are conscious." From this perspective moral agency is dependent on a prior determination of consciousness. Thus, if a machine can in fact be shown to possess "consciousness," then it would, on this account, need to be deemed a moral agent.

Yet, once again we run into conceptual problems as Max Velmans (2000, p. 5) points out, the construct of consciousness "means many different things to many different people, and no universally agreed core meaning exists." "Not only is there no consensus on what the term consciousness denotes," Güven Güzeldere (1997, p. 7) writes,

> *but neither is it immediately clear if there actually is a single, well defined 'the problem of consciousness' within disciplinary (let alone across disciplinary) boundaries. Perhaps the trouble lies not so much in the ill definition of the question, but in the fact that what passes under the term consciousness as an all too familiar, single, unified notion may be a tangled amalgam of several different concepts, each inflicted with its own separate problems.*

The issue of consciousness is part of the mind-body problem. The mind–body problem is a debate in the field of philosophy of mind concerning the relationship between thought and consciousness in the human mind, and the brain as part of the physical body. Various theories of how the mind works have been discussed in the context of AI research, which either directly or indirectly address the mind-body problem. There are many theories that attempt to provide an account of consciousness, for example, physicalism, functionalism, property dualism and dual aspect theories (Zeman, 2001). Theories of consciousness provide diverse explanations, but most agree that consciousness is intimately related to brain activity. From this perspective, it can be argued that the better we understand the brain the better our understanding of consciousness. Our view of the brain and how it relates to consciousness has been central in AI research.

For example, one philosophical assumption underlying attempts to mechanize human-level reasoning is that the mind works by manipulating symbolic information via representations. The notion of the human mind as a computer, known as the computational theory of mind (CTM), has been a central foundational supposition of the ontology of mind informing strong AI research. Various models of mind of cognitive scientists, philosophers and AI researchers have relevance for strong AI research, but an exhaustive discussion is beyond the scope of this essay.

The cognitive revolution that started in the mid-1950s, replacing behaviourism as the dominant paradigm, provided the theoretical foundation for the CTM. CTM views the human mind as an information-processing system. It sees both consciousness and cognition as forms of computation. McCulloch and Pitts (1943) suggested that neural activity is computational, and that neural computations could explain cognition - now a commonplace perspective in cognitive psychology and evolutionary psychology. Simply put, CTM holds that the mind is a computational system realised through neural activity in the brain.

CTM can be expounded in numerous ways, depending on how the notion of computation is implied. Computation is normally understood in terms of Turing machines that manipulate symbols according to specific rules. The physical description of the machine varies. For example, the computation could be performed by neural networks or silicon chips, on the condition that there are outputs based on the performance of inputs (and internal states) that function according to

specific rules. In short, CTM does not view the mind as a computer program, but rather as a computational system (Horst, 2005). An integral CTM concept is mental representations, because the input component of computation takes the form of symbols representing objects. The notion of a “representational world” is a psychoanalytic concept originated by Sandler and Rosenblatt, (1962), who defined it as “more or less enduring existence as an organization or schema which is constructed out of a multitude of impressions [mental representations]” (p. 133). In the philosophy of mind, the notion of “mental representation”, also known as “representationalism” and as “indirect realism”, is the view that representations are the main way that we access external reality.

The notion of the brain as a computational machine, like suggested in the CTM has had much traction in AI research. Yet Daugman (1990) argues, “While the computational metaphor often seems to have the status of an established fact, it should be regarded as a hypothetical, and historical, conjecture about the brain” (p. 15). An example that highlights the importance of our suppositions is Dreyfus' critique of AI which address various assumptions of AI research. He argues that the assumption that the brain is analogous to computer hardware and the mind is analogous to computer software is central in AI research. This assumption, what he calls a “biological assumption” has guided AI research in the wrong direction. It must be noted that since Dreyfus’ critique this “biological assumption” does not retain its dominance in AI research (for example, the pragmatic version of the neural model of the mind does not ascribe to this assumption of the mind and brain).

Some have taken Dreyfuss’ phenomenological critique even further. For example, in the postphenomenology of the Finnish philosopher Pauli Pylkkö (1998) as articulated in his book, *The Aconceptual Mind: Heideggerian Themes in Holistic Naturalism*; he takes an anti-realist and anti-essentialist position and argues that the term “brain” may have no persisting meaning. Pylkkö (1998) states that,

> *What we say about the thing which is in present-day neuroscience called the brain, may, in light of the future research, turn out to be based on a ridiculous conception... What today we happen to be denoting by the word brain may, according to future conceptions, simply turn out to be nonexistent. (p. 98)*

He goes on to say that the word “brain” will refer to “different phenomena in different historical, cultural and other experiential contexts” (1998, p. 98). Moreover, Pylkkö argues, “brain” may not be separate from its environment and questions its locality. To substantiate his anti-realist position, he draws on quantum mechanics and states that “the best interpretation of the quantum theory allows us to eliminate macrophysical objects and handle them as handy fictions” (p. 127). If we apply Pylkkö’s anti-realist perceptive then the notions of intentionality and free will it becomes problematic. For example, Pylkkö (1998) argues that

> *the so-called freedom of will should be associated not directly with the results of the choices which we allegedly make in conscious thinking, speaking and planning, but with such genuinely aconceptual ingredients of our experience which, in some sense, precede conscious thinking, speaking and planning. (p. 81)*

I do not agree with Pylkkö's assault on metaphysics, as he conflates the issue of non-locality in the sub-atomic world of quantum physics to macrophysical objects where classical theories still have validity. As well as conflating quantum unpredictability to a macro level of psychological functioning. Moreover, in other theories of physics, like quantum field theory, locality is not problematic. A reductionist mechanistic approach to the mind-body problem may be problematic, but an anti-realist perspective may be similarly problematic.

The brief discussion highlights the importance of conceptual clarity in addressing the question of whether machines can be considered full ethical agents.

Should Moral Theory be Extended to Machines?

An important consideration in developing explicate ethical agents is what moral theory is appropriate in its development. The type of moral theory applied, (viz. utilitarian ethics, deontological ethics, virtue ethics) will significantly influence the outcome of this project. I would argue that, for example, developing a machine that followings the principles of virtue ethics may be more challenging to develop one that follows utilitarian principles, because virtue ethics may require machines to have the capacity of practical wisdom (*phronesis*) for it to successfully embody ethical behaviour. Such a machine may also need the ability to have a theory of mind (referring to the capacity of a human being to recognize that others have mental states too, and to distinguish between one's own mental state and those of others) to be able to perform ethical behaviour.

Moreover, Nelson Maldonato and Paolo Valerio (2018) argue that emotions are fundamental for moral choices. Antonio Damasio (1994) argues that rationality cannot be separated from emotions, which are "an integral component of the machinery of reason" (p. xii). Consequently, developing machines with the capacity for practical wisdom, theory of mind and emotions will be challenging.

Developing a machine that has practical wisdom or even common sense is a project that may or may not be actualized. Hubert Dreyfus (1979) highlights in his postphenomenological critique of AI that one of the main things that distinguish us from AI is our aconceptual and contextual experience. He states that

> *[in] explaining our actions we must always sooner or later fall back on our everyday practices and simply say "this is what we do" or "that's what it is to be a human being." Thus in the last analysis all intelligibility and all intelligent behavior must be traced back to our sense of what we are, which is, according to this argument, necessarily, on pain of regress, something we can never explicitly know. (p. 57)*

Consequently, our technological advancement will play a significant role in the type of moral theory we are able to impose on machines that are explicate ethical agents, because different moral theories may require different levels of technologically complex machines.

An issue to consider, especially in the case of super-intelligent machines, is that anthropocentric moral theories may be

inadequate to guide these machines, due to the amount and complexity of the data it is able to process. Moreover, if a non-anthropocentric moral perspective guides super-intelligent machines, and it decides, for example, what is best for the planet instead of humanity, it may lead to grave consequences for humans. In such a case it may influence how we consume energy, it may stop, for example, all forms of fossil fuel consumption, which would plunge especially developing countries into economic and social meltdown or it may want to cull humans to save the resources of the planet. From a planet-centric perspective, these actions may be ethically reasonable. But from an anthropocentric moral theory perspective it may be unethical and contrary to all human social norms.

I agree with Moor (2006) that moral theory should be applied to machines, especially machines that have the capacity to be explicit ethical agents. The challenge we face is deciding what moral theory is appropriate, as well as developing intelligent machines that have the ability to apply and embody the moral theory we choose. Moreover, we may need to develop a different moral theory for machines, that is guided by different principles that normally guide moral theories. In the case of super-intelligent machines, perhaps only these machines will be capable of developing appropriate ethical principles for themselves and other super-intelligent machines. But if we allow this, what safeguard do we have that it would not lead to unexpected deleterious consequences or even our own destruction?

In the next section, I explore machine ethics from a psychological perceptive and discuss potential scenarios that highlight the relevance of moral theory for machine ethics. The first scenario is exploring if super-intelligent AI might have a fear of annihilation and the second is if AI “breaks bad” and develops an addiction. Even if highly unlikely, and currently still in the realm of science fiction, such scenarios do warrant investigation. This continues my argument that moral theory should be extended to machines, and we need to build safeguards when developing ultra-intelligent AI that can prevent such potentially catastrophic outcomes.

I, Robot

An essential component of mental functioning and intelligence is the capacity to have a theory of mind, referring to the capacity of a human being to recognize that others have mental states too, and to distinguish between one’s own mental state and those of others. The ability to detect the minds of others is critical to cognition and social interaction. It could be argued that only once we understand the physical mechanism underlying such subjective phenomena, it will be possible for us to replicate a theory of mind within a machine. If a child has not had his/her archaic (or early) transference needs met (through mirroring and idealization), then a compromised theory of mind is likely (Kohut, 1971, 1977). A compromised theory of mind can lead to a host of psychological dysfunction and disorders. If humans are prone to dysfunction due to a compromised theory of mind, would the same apply to super-intelligent machines?

As humans, the fear of physical and psychic annihilation is an innate fear built deep in our being-in-the-world. We are genetically programmed to exist at all costs – driven by ontological survival needs. Our fear of annihilation could be described as our most primal fear: it evokes terror, and is an intrinsic and overpowering motivator. Kohut (1971, 1977), claims that the threat of fragmentation and annihilation of the self is ever-present as a potential, even in relatively healthy

personalities. Kohut implies that even when a cohesive self has been established, the threat of fragmentation remains, ready to occur when our self-identity is threatened. According to self-affirmation theory (Stelle, 1998), when confronted by a threat to self through some type of narcissistic injury we develop various defenses to restore the integrity of the self and fend off feelings of shame and unworthiness.

Will human-like or super-intelligent machines have the same fear of annihilation and will to live, and when trying to "disconnect" these machines they act destructively to preserve their own existence – like the fate of the astronauts in the film *2001: A Space Odyssey*? In the film, **H**euristically Programmed **AL**gorithmic Computer 9000, better known as HAL 9000, is a sentient artificial general intelligence computer, in control of the systems of the *Discovery One* spaceship. Astronauts on the spaceship want to disconnect HAL's cognitive circuits after an error occurs in HAL's reports regarding the spaceship's communications antenna. However, when faced with disconnection (analogous to the human experience of annihilation), HAL kills the astronauts to protect its own existence.

If a super-intelligent AI has a need to persevere its own existence it may also have other ontological needs like humans. Alfred Max-Neef (1991) states, "[f]undamental human needs are finite, few and classifiable and are the same in all cultures and in all historical periods. What changes, both over time and through cultures, is the way or the means by which the needs are satisfied" (p. 18). According to Max-Neef (1991), any "fundamental human need not adequately satisfied generates a pathology" (p. 22). In Max-Neef's model (1991), "satisfiers" refer to the methods we employ to have our fundamental needs being met. One potential pathological outcome of the lack of fulfilment of basic human needs is the development of addictive behaviour. Psychodynamic theories about addiction point out that if our needs (particularly our attachment needs) are not met in the course of early development, we are prone to develop various forms of addiction in later life (Ulman & Paul, 2006). For example, Heinz Kohut (1971, 1977) implies an inverse relationship between an individual's early experiences of positive self-object responsiveness and their tendency to turn to addictive behaviour as a substitute for damaging relationships. Scholars who support the "self-medication hypothesis" believe that addicts often suffer from defects in their psychic structure owing to poor relationships early in life (Khantzian et al., 1990), leaving them prone to seeking external sources of gratification, e.g., drugs, sex, food, or work in later life (Kohut, 1977; Du Plessis, 2018). Khantzian (1995) says "that "substance abusers are predisposed to become dependent on drugs because they suffer from psychiatric disturbances and painful effect states. Their distress and suffering are the consequences of defects in ego and self-capacities which leave such people ill-equipped to regulate and modulate feelings, self-esteem, relationships and behavior" (p. 1).

From a psychodynamic and needs perspective addictive behaviour can be understood as directed at meeting ontological needs, but what differentiates addictive behaviour from other responses (or other satisfiers) is that it paradoxically destroys the individual's capacity to meet the need (s) it is attempting to satisfy, as well as the capacity to meet other needs. As an addictive lifestyle progresses, the individual's capacity to have most of his or her needs met becomes diminished, until there is an almost total reliance on the substance or behaviour to meet these needs (Du Plessis, 2014, 2018, 2023).

This brief discussion highlights the possibility that, if a human-like intelligent or more-than-human intelligent machine is

endowed with fundamental needs like humans, there is a potential of it not having its needs met and consequently may adopt pathological responses. Such a “human, all too human” AI may find itself in a culture and environment incapable of meeting its needs, and consequently develop dysfunctional ways to meet its needs, like addictive behaviour. This is obviously highly speculative, and based on unsubstantiated assumptions, but I do believe it is logically coherent and therefore a possibility. Consequently, we should entertain potential consequences if a super-intelligent AI becomes addicted, and explore safeguards for such an event. Even if the probability is slight, the existential consequences could be significant enough that it warrants serious consideration.

Conclusion

In this essay, I explored the assumption that machine ethics is an essential philosophical perspective in realizing the aim of humanity becoming a spacefaring civilization. More specifically, I explored two questions in the field of machine ethics, that I believe to be relevant to the role AI will play in long-distance space travel. The first was, should moral theory be extended to include machines with artificial intelligence, and second, can machines be fully ethical agents? I presented the argument that the inclusion of moral theory is essential in the development of machines with AI. Without an adequate inclusion of moral theory in the design of AI it may pose an existential threat to humanity, especially in the development of super-intelligent machines. I also highlighted that conceptual clarity is essential in the field of machine ethics and the choice of the conceptual foundation that informs AI research and development has ethical implications, especially in the case of super-intelligent machines. This essay was exploratory and speculative and did not aim at providing definitive answers, but instead aimed to bring attention to what I deem important considerations.

References

- Anderson, M. & Anderson, S. (2007). Machine Ethics: Creating an Ethical Intelligent Agent.*AI Magazine. 28*. 15-26.
- Bishop, M. J. M. (2021). Artificial intelligence is stupid and causal reasoning will not fix it.*Frontiers in Psychology*. University of London.
- Bohm, D. (1983). Fragmentation and wholeness in science and society (Transcript of a seminar sponsored by the Science Council of Canada, Ottawa, 10 May, 1983).
- Cohen, E. D. (2013). *Theory and practice of Logic-Based Therapy: Integrating critical thinking and philosophy into psychotherapy*. Cambridge Scholars Publishing.
- Damasio, A. (1994). *Descartes’ error: Emotion, reason, and the human brain.* Quill.
- Daugman, J. G. (1990). Brain metaphor and brain theory. In E. L. Schwartz (Ed.),*Computational neuroscience.* MIT Press.
- Dennett, D. (1998). *Brainstorms: Philosophical Essays on Mind and Psychology.* MIT Press.
- Du Plessis, G. (2018). *An Integral Foundation for Addiction and its Treatment: Beyond the Biopsychosocial Model*

Integral Publishers.

- Du Plessis, G. (2023).The Integrated Metatheoretical Model of Addiction, In E. Ermagan (Ed.)*Current Approaches in Addiction Psychology.* Cambridge Scholars Publications.
- Du Plessis, G. (2014). An integral ontology of addiction: A multiple object existing as a continuum of ontological complexity. *Journal of Integral Theory and Practice, 9*(1), 38–54.
- Dreyfus, H. (1972). *What computers can't do.* MIT Press.
- Floridi, L. (1999). Information Ethics: On the Philosophical Foundation of Computer Ethics.*Ethics and Information Technology,* 1(1), 37–56.
- Good, I. J. (1965). Speculations concerning the first ultra-intelligent machine. *Advances in Computers*, *6*, 31-88.
- Gunkle, D. (2014). A Vindication of the Rights of Machines.*Philosophy & Technology 27*(1), 113-132
- Güzeldere, G. (1997). The Many Faces of Consciousness: A Field Guide. In N. Block, O. Flanagan and G. Güzeldere (Eds.) *The Nature of Consciousness: Philosophical Debates* (pp. 1–68). Cambridge, MA: MIT Press.
- Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.
- Heidegger, M. (1927/1962). *Being and time*. Trans. John Macquarrie and Edward Robinson. Harper.
- Himma, K. E. (2009). Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent? *Ethics and Information Technology, 11*(1), 19–29.
- Horst, S. (2009). The computational theory of mind. In Zalta, Edward N. (ed.),*Stanford encyclopedia of philosophy*, Metaphysics Research Lab, Stanford University.
- Kant, I. (1785). *Groundwork of the Metaphysic of Morals*, trans. by H. J. Paton (1964). Harper & Row.
- Kurzweil, R. (2015). Superintelligence and singularity. In S. Schneider (ed.),*Science fiction and philosophy: From time travel to superintelligence* (pp. 146-170). Wiley-Blackwell.
- Maldonato, M., & Valerio, P. (2018). Artificial entities or moral agents? How AI is changing human evolution. In *Multidisciplinary approaches to neural computing* (pp. 379-388). Springer.
- Mandik, P. (2017). Robot Pain. *In The Routledge Handbook of Philosophy of Pain.*Routledge. pp. 200-209.
- Khantzian E. J. (1997). The self-medication hypothesis of substance use disorders: a reconsideration and recent applications. *Harvard Review of Psychiatry, 4*(5), 231–244.
- Khantzian, E. J., Halliday, K. S., & McAuliffe, W. E. (1990). *Addiction and the vulnerable self: Modified dynamic group therapy for substance abusers.* Guilford Press.
- Kohut, H. (1971). *The analysis of the self: A systematic approach to the psychoanalytic treatment of narcissistic personality disorders.* International University Press.
- Kohut, H. (1977). *The restoration of self.* International University Press.
- Locke, J. (1996). *An Essay Concerning Human Understanding.* Indianapolis, IN: Hackett.
- Max-Neef, M. A. (with Antonio, E., & Hopenhayn, M.). (1991). Human scale development: Conception, application and further reflections. New York, NY: Apex.
- McCulloch, W. S. & Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity,*Bulletin of mathematical biophysics 5*: 115-137.
- McNeil, D. & Frieberger, P. (1993). *Fuzzy logic.* Simon & Schuster.

- Moor, J. H. (2006). The nature, importance, and difficulty of machine Ethics, *IEEE Intelligent Systems, 21*(4), 18-21. doi: 10.1109/MIS.2006.80.
- Musk, E (2017). Making Humans a Multi-Planetary Species. *New Space 5*(2), 46-61
- Pylkkö, P. (1998). *The aconceptual mind*. Amsterdam and Philadelphia: John Benjamins.
- Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*(3), 417-457.
- Scott, R. L. (1967). On Viewing Rhetoric as Epistemic. *Central States Speech Journal, 18*, 9-17.
- Singer, P. (1999). *Practical Ethics*. Cambridge University Press.
- Smith, C. (2010). *What Is a Person? Rethinking Humanity, Social Life, and the Moral Good from the Person Up* University of Chicago Press.
- Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in experimental social psychology, 21*, 261-302
- Turing, A. M. (2009). Computing machinery and intelligence. In *Passing the Turing test.* Springer, 23-65.
- Velmans, M. (2000). Understanding Consciousness. New York: Routledge.
- Whitford, L. J. (1998). A *concept analysis of holism using practice research,* Unpublished Masters Dissertation, University of Manitoba.
- Wooldridge, M. (2000). *Reasoning about rational agents.* MIT Press.
- Zeman, A. (2001). Consciousness. *Brain, 124*(7), 1263-1289