**Scientia*et*Fides** 13(1)/2025 ISSN 2300-7648 (print) / ISSN 2353-5636 (online)

Received: November 22, 2024. Accepted: January 7, 2025 DOI: http://dx.doi.org/10.12775/SetF.2025.009

# Skynet Meets Planet of the Snakes. Removing metaphysical Impediments to Rogue AI

**ANDY MULLINS**
University of Notre Dame Australia
apjm.vic@gmail.com
ORCID: 0000-0003-1540-3796

**Abstract:** This paper advances the view that non-rational AI could indeed constitute a significant threat to humankind and that the best way to avert this threat is ethical formation of those involved in AI development. In the light of a novel thought experiment, and by applying an understanding of rationality based on a Thomistic Metaphysics of Participation (TMP) wherein rationality and self-awareness do not emerge, this paper finds no metaphysical impediment to AI posing such a threat in the future. This approach serves to highlight the immediate need for action. The most effective safeguard against a Skynet scenario appears to be ethical formation in justice for the policy and decision makers, and AI developers.

**Keywords:** Skynet, artificial intelligence, rationality, Aquinas, participation, justice.

"If we continue to develop our technology without wisdom or prudence,
our servant may prove to be our executioner."
General Omar N. Bradley

## Introduction

The exponential increase use of AI on the battlefields of the world excludes all trite discussion. Pope Francis' 2024 message for the World Day of Peace, called for the imperative necessity of ensuring "adequate, meaningful and consistent human oversight of weapon systems." So with that solemn caveat let us reflect on the Skynet scenario in which unshackled AI makes war on humans. Could that be possible? And if so how could we avoid it?

Some think it impossible, arguing that self-awareness is a bridge too far for technology, that machines do what they are programmed to do, and that fears are completely overhyped. Wang insists that machines cannot be a threat as they lack self-awareness and self-conscious emotions, and therefore any "internal driving force of dominating" (Wang 2023, 73). Others consider that "the current threat of AI is vastly overstated and the technological singularity remains a distant theoretical danger" (Tredinnick and Laybats 2023).

Others are not so sure. Liu notes the unknown effects of decoupling of intelligence from emotion and from consciousness (Liu 2023, 819–820), and Chatila et al (2018, 88) highlight the current lack of even a simple understanding of how machines could "understand their environment, to be cognizant of what they do, to take appropriate and timely initiatives, to learn from their own experience and to show that they know that they have learned…"

And for others again, AI is a great concern, either in the hands of man, or as a rogue entity. Vladimir Putin proclaimed that the country to lead in development of AI will "become the ruler of the world" (Scharre 2019). Geoffrey Hinton (2012), founder of the *neural network* approach that powers current AI, has warned last year "There's a serious danger that

we'll get things smarter than us fairly soon and that these things might get bad motives and take control" (Allyn 2023).

A 2015 letter by Stephen Hawking, Elon Musk and others, eulogised AI, but in 2023 a second letter signed by 1000 industry leaders and researchers, Musk among them, wrote of the "profound risks to society and humanity" that will accrue from careless development of "AI systems with human-competitive intelligence".

It urged a six month pause on advanced AI development: "AI labs and independent experts should use this pause to jointly develop and implement a set of shared safety protocols for advanced AI design and development that are rigorously audited and overseen by independent outside experts. These protocols should ensure that systems adhering to them are safe beyond a reasonable doubt. Shortly later a further formal statement from OpenAI, Google DeepMind, Anthropic and other labs and companies spoke of AI causing an "extinction event" (Vice, Motherboard blog). Yet there is a significant disconnect between such concerns and pragmatic commercial interests. Ironically, twelve months on, it appears not only has there been no pause but development has accelerated. (*Wired* 28 September 2023)

Current developments highlight further concerns. In a 2023 USAF simulation an autonomous drone "neutralised" its handler for impeding its mission. The simulation has now been downplayed. (*Reuters* 9 June 2023) In 2023 also, a researcher affiliated with Google Deepmind co-authored a paper concluding that a world-ending catastrophe was "likely" if a rogue AI were to come up with unintended strategies to achieve a given goal, including "[eliminating] potential threats" and "[using] all available energy." (*Vice* 1 June 2023)

There is polarised disagreement also about the very meaning of the term intelligence in relation to a machine. Fjelland (2020) soberly points out that the development of artificial general intelligence is "in principle impossible". Another writes, "although AI may exhibit behaviours indicative of self-awareness, the subjective experience of self-awareness remains a uniquely human trait […]." (Namestiuk 2023, 44) But Dutt et al. (2020, 971) suggest that self-awareness of a machine is constituted by "infering its own state in relation to environment", and Gonzalez-

Jimenez urges 'the integration of self-aware robots in society' (Gonzalez-Jimenez 2018, 18). Others warn against the catastrophic risks in AI as it becomes "more intelligent than we are." (Hendrycks, et al. 2023) And Hunt attributes agency to superintelligent AI which will soon be "able to run circles around programmers and any other human by manipulating humans to do its will." (Hunt 2023)

In summary perceptions of the threat posed by AI are wildly varying and this seems related to the metaphysical conundrum of whether a machine can in the future manifest self-awareness and agency. But such disagreement paralyses the development of public policy. Meanwhile, AI development accelerates without effective guidelines. There remains no consensus whatsoever about the need to mitigate fears, and what such mitigation might look like.

## 1. Methods

By thought experiment I argue that non-rational AI, operating as a predatory animal, could indeed constitute a major global threat. I then argue, utilizing a stringent understanding of participated, non-emergent rationality, that the notion of a rational self-aware Skynet is not possible in any case. Yet the combination of the thought experiment with the metaphysical reflection therefore contributes substantially to the view that a Skynet scenario may be closer than many believe. And by removing the rationality conditions that constitute the metaphysical impediment to a Skynet scenario but by showing that AI can nevertheless present a global threat I clarify certain challenges for safe development of AI. I place the onus back on virtuous development rather than intrinsic programming safeguards. Global protocols are required to ensure that policy and decision makers be formed in justice, in the habit of overarching respect for other persons.

This paper therefore argues that the current debates about whether or not the development of AI brings significant dangers are themselves dangerous. AI will not eventually become sentient and compete with humankind centuries in the future; it need satisfy only a far lower

threshold of autonomous behaviours to become a threat. By dismissing the mistaken notion that sentient AI with agency will become a big problem, we can focus more effectively on the immediate reality.

### 1.1. A thought journey

In our thought experiment, we visit the Planet of the Snakes, a place, we are told where snakes have evolved animal behaviours documented across the animal kingdom. In describing their behaviours, I draw on established knowledge about the capacity for animals to communicate and cooperate with each other, to recognise intentions and mental states, to monitor friendships and animosities, to collectively organise, and to act in pursuit of tangible goals (cf. Cheney 2011; Whitehead 1997). Let the tour guide take up the story:

> On our galactic tour we have now arrived at the Planet of the Snakes which you can see below you. Sadly no human life now survives in this world. The last died agonising deaths several millennia ago. *Reptilia squamata,* the snakes you are familiar with on Terra, are but benign antecedents compared with the serpents inhabiting, or we could say, 'ruling' this planet. Terra snakes are solitary and asocial, living in miserable holes and rotten logs to avoid harassment. But as you know from the pre-flight briefing, the snakes on this planet have no natural predators and have evolved complex social structures and hierarchies. They have eliminated all competition and colonised the planet. Their abodes are complex burrows and mounds, at times constructed by lizards and insects, pressed into service and later sacrificed as food. A hierarchy of power is established in each nest. Younger snakes have worker roles and female snakes remain in permanent nurseries guarding eggs. And they exhibite a quasi-compassion for their own. The diarist noted,
>
> > I have seen many snakes congregate around the body of a snake we had managed to kill. To observe these "ceremonials" was to take one's life in one's hands, because, like a mob of kangaroos, they had lookouts positioned, and if the onlooker be sighted, his cause would be lost.
>
> In this snake society their castes and defence mechanisms are reminiscent of the colonies of wasps or ants. We still do not understand their advanced forms of communication, made up of visual and vocal cues and chemical secretions.

> And their predatory instincts have evolved to the point that they could identify particular humans. As you will know, human survivors were driven into caves where single entrances could be guarded more or less effectively. Yet they were no match for the snakes who had evolved cooperative foraging and hunting practices that were highly organised: some snakes would flush out prey, while others positioned themselves in ambush. Listen to the diarist,
>
> > We lost many who, fleeing from apparently a chance meeting with snakes, in their escape were channelled into dead end canyons, writhing with serpents.
>
> Please note that our visit to this planet is brief. If you are in descent party you have received instruction for use of your rocket pack to escape difficult situations. Do keep your wits about you and tread carefully. Extraction after injury or death will not be possible. This was clearly stated on the misadventure waiver that you all signed.

## 1.2. Reflections on limitations

The limitations of this study appear to be twofold.

First, the above thought experiment, while eye catching, may seem extravagant and fail to convince because of the anthropomorphic application of certain principles of rational psychology, such as the distinction between sense appetite and rationality. If those presuppositions are accepted, the thought experiment retains validity. We see that the Planet of the Snakes is dominated by highly effective predators and observe that the "intelligence" of the snakes, as for many animal predators in our own world, need only be motivated by physical sensible appetites. So too, I hypothesise that a Skynet scenario could also be triggered by physical sensor-dependent conditionings. While such conditionings do not constitute rationality, they do mimic the sensible appetites of learned responses to pleasure and fear that we see in predators; responses that are unguided by conscience (with conscience understood as an internal capacity for grasping the correlation of action scenarios against moral truth convictions).

Second, it is beyond the scope of this paper to enter into the great variety of views on rationality. However, given the pointlessness of

any exercise founded on a reductive vision of rationality, I opt for a metaphysical view of rationality in order to defend the spiritual nature of man. Karol Wojtyla wrote of the necessity of metaphysics if we are to be human:

> Metaphysics should not be seen as an alternative to anthropology, since it is metaphysics which makes it possible to ground the concept of personal dignity in virtue of their spiritual nature. John Paul II, 83.

Within hylomorphic philosophy of mind there are two broad approaches to account for rationality: one looks to Aristotelian formal causality as the key focus in accounting for rationality. This approach is found for example in the work of Feser (2005), Madden (2013), Jaworski (2016), De Haan (2018); and the second offers an elaborated understanding of Thomistic participation in being, a view championed for example by Norris Clarke, Koterski (1992) and Wippel (2000).

I propose to utilise this second approach, an understanding of rationality founded on a Thomistic metaphysics of participation (TMP). Thus measuring Skynet against a strict test for rationality we offer a coherent critique of assumptions about cognitive thresholds and self-awareness.

I elaborate the characteristics of rationality looking first at participation in being and then at rationality itself. This will provide a lens through which we can assess the potential of machines to acquire rationality. I will suggest that rationality is a state of nature disposing to certain behaviours, rather than simply a description of those behaviours. And therefore it may be argued that the notion of rational machines has no meaning.

### 1.3. A metaphysics of participation

Within Aquinas' metaphysics, being and rationality are obtained by participation in God's being. As such they are bestowed spiritual realities and cannot emerge from matter.

The notion of participation is crucial to Aquinas and he utilises the term in various ways. He identifies three modes of participation: logical

(species within genus); when a subject in accident and matter participates in form; and a third mode in which an effect is said to participate in its own cause, such as a finite being in *esse* (EHB, Ch2). It is the third mode that interests us here. Fabro taught that Aquinas' dialectic of participation was "the hermeneutic key of the originality of Thomism" (1969, xxxiii). Aquinas presents a framework in which all beings derive their existence from a singular source, with form participating *in esse subsistens*. (Mullins 2022b, 182)

The necessity of a metaphysics of participation becomes evident when considering the argument from contingency. This argument hinges on the recognition that "existence is something other than the essence or quiddity." (DE, III, 77) Embodied rational subjects enter and exit embodied existence, unlike matter which perdures: before conception, the person is nonexistent, and after death, the person has departed, though some physical evidence remains. The soul is not originating from matter yet animating a contingent substance. It is an intrinsic principle of existence and rationality, bestowed from outside matter, from "the first cause" (DE, III.80) This formal source of being, life, and function, is thus also the principle of personhood.

In contrast with the teaching of Aristotle of soul as the principle of activities, Aquinas argued that the soul must be a participating principle of existence (ST, 1.3.4; Wippel 2003, 8). He presented form not only as a principle of function or unity, but as a participation *in esse subsistens* (Fabro 1970, 71–72). This highlights a major point of difference with Thomistic philosophies of mind that do not draw attention to participation *in actus essendi*. Thus, the existence of an intellectual being is directly dependent on an Ultimate Source, imbuing the subject and its rationality with profound dignity.

This understanding of participation as the fundamental relationship between creatures and their Creator is central to Christian philosophy. By attributing being to an Ultimate Source (to use the words of Norris Clarke) TMP grounds existence, unity, and function in being. "All other beings that are not their own being but have being by participation must proceed from that one thing," (DP, 3.5) and elsewhere, "That which has existence but is not existence, is a being by participation. (ST 1, q. 3.4)

Aquinas presented *esse* as *actus essendi* in contrast to *existentia* of Augustinianism and of rationalism (Fabro 1974, 449). Bazan notes, "(this) provides the ultimate foundation of his anthropology, namely the real distinction between *esse* and *essentia* and the philosophical theory of creation as causation of the finite act of being (*esse*) by an Infinite Being (*Esse subsistens*)" (1997, 114)." To explain the real distinction between essence and existence, he applied the notion of act and potency to being, giving primacy to the act of being (Rziha 2009, 7): "It is from the concept of *esse* as ground-laying first act that Thomas develops his own notion of participation and his entire metaphysics." (Fabro 1974, 463; cf. 449)

In contrast to Aristotelian hylomorphic accounts, the Platonic tradition, and Thomistic approaches that neglect participation, the emphasis on participation establishes existence as prior through explicit metaphysical demonstration. Without respecting existence as prior, the objectivity of reality and causality are at risk of being lost, undermining the integral convertibility of truth and being at their transcendent foundations by severing their mutual participation in *ipsum esse subsistens*.

So it is seen that rationality depends not on the rational recipient but on the bestower, the Ultimate Source. Neither biological entities, nor machines, regardless of complexity or evolutionary excellence, can become a rational being. This is not an inherent capacity.

### 1.4. Human beings are rational by participation

Rationality itself is a participation, intrinsic to, and inseparable from, personhood, understood as "a subsistent individual in a rational nature". (ST, I, q. 29, a. 3). Within Aquinas' view, it would be a metaphysical impossibility that machines could become rational. In *Contra Gentiles*, Aquinas emphasizes the operation of understanding as an activity "completely surpassing the range of bodily things." (SCG, 2.86.7) And he notes in *De Veritate* how being also underpins knowing. (QDdV, 1.1. 5)

Aristotle's explanation for life, that the animal soul emerges from matter, cannot suffice for human beings. As human intellectual capacities surpass the limitations of matter, enabling them to comprehend non-material realities and exercise free will, Aquinas argued that purely

physical entities cannot account for intellectual life, as the greater cannot be produced by the lesser: "no corporeal power can produce the intellective soul" (SCG, 2.86.7).

This principle of being is the principle of rationality, "the human being understands through the soul." (ST, I, q. 75, a. 2) Rationality should not be seen as a ghostly phenomenon outside the world of matter, but rather as an operation that is transcendent yet within matter itself by virtue of a participated power.

All knowledge originates from the senses in the apprehension of a phantasm or image in our minds. Drawing on Aristotle Aquinas explains that through the light of our participated active intellect, attuned to being and act, we comprehend the object in the passive intellect, perceiving its existence and substantial form. Note that Aquinas underscores operations rather than structures, and the independence of understanding from the body itself.

All mental life is supported at the neural level, but not reducible to the biophysical. By analogy, when one whistles the physical action is distinct from one's appreciation of one's own whistling. So understanding depends upon yet transcends physics and biophysics. The physical is intrinsic to and inseparable from the transcendent understanding.

But the physical is *always* involved. In human embodied existence the act of understanding requires the material mediation by the physical for *every* operation of understanding and willing. In this embodied life understanding and willing occur in a transcendent domain intrinsically dependent on the physical.

Rationality, while mediated by physical processes, is not reducible to them. The active intellect, consisting of immaterial and intelligible species, facilitates the actual intelligibility of phantasms, participating in what is termed "Divine light" (ST, I, q. 89, a. 1 and I, q. 84, a. 6). In the same way that the soul is present, neither as agent nor as substance, these immaterial species are present *by participation*. They are within matter but transcending matter. This underscores the idea that the intellect is a participated power, not inherently its own but belonging to another entirely (Fabro 1974, 454). Aquinas held the view that goodness and truth participate in being, and humans possess both being and rationality

through participation. This participation in being enables non-material perfections, such as the capacity to know and will, to be essential properties of human nature. Rationality, thus, is not merely a process but an intrinsic aspect of human nature, integral to the essence of humanity.

It should be noted that proponents of hylomorphism hold diverse views on immateriality. For instance, Robert Pasnau (1998) has argued that the hylomorphic explanation conflates ontological and representational immateriality. TMP offers an alternative view. (Mullins 2022a) "Immateriality of thought" rather refers to operations not structures or physical representations present within ensouled matter (SCG, 2, 90, 4). These operations are mediated by the physical realm but cannot be reduced to physical processes. Although "the human being understands through the soul." (ST, $I^a$, q. 75, a. 2.) the embodied, ensouled person is the agent, not the soul.

### 1.5. And what of teleology and personal fulfilment?

The operations of rationality make possible human fulfilment in truth and love. Non-reductive physicalist accounts, including hypotheses of machine intelligence, cannot account for such fulfilment as essential, because they offer no basis to prioritise the operations of rationality over other activities… all are absolutely and inherently material. (Mullins 2022b) Within TMP however, rationality is an essential quality prior to all other features. TMP provides a coherent account of the capacity for human beings to grasp truth and universal concepts, to make love choices based on these truths, prioritising therefore persons capable of reciprocal love. And what of choice, the capacity to love? Developed from a communitarian Thomistic perspective, Walker argues that loving relationships are integral to rationality; by their very nature human persons are fulfilled in personal loving relationships. (Walker 2004)

As the intellect is a spiritual power, so too is the rational will. (SCG, 90,4) Matter alone cannot explain consciousness, self-knowledge, knowledge of external realities, or choices utilising that knowledge. Note that operations are acts: "The powers of the soul are known through their acts" (ST II, q. 77, a. 7). It is the acts that are immaterial, not the medium.

Hence the capacity to love is an immaterial operation, only possible through participation in "Divine light". (Mullins 2022b)

It is the self-diffusion inherent in the Primary act of being that gives rise to man's capacity for self-giving love, as a participation in being itself (DP, 2.1). Aquinas viewed *ens* as "not simply *essentia* or *esse*; rather it is the selfgivenness in act of their synthesis (Fabro 1966, 403). Thus, man's ability for and fulfillment in self-giving love can be traced back to participation *in esse subsistens*, framing *esse* as the pure act of self-donation (Schindler 2005, 19). Hence man is fulfilled in giving of himself at the transcendent level of the person, not at a material level where one is diminished in the giving.

### 1.6. Considerations of personhood and self-awareness

If the highest operation of rationality involves loving wisely, loving other persons primarily, then the first choice should be to love God in acknowledgement that he is the source of all other goods. Norris Clarke held "that man could not be oriented toward union with God by the innate drive of his spirit unless there were some kind of profound ontological affinity or similitude." (Norris Clarke 1981, 516) Such similitude, as it is founded on participation, is impossible for AI.

Being, personhood, rationality, and love are inherently interconnected concepts within Thomistic metaphysics, Aquinas views personhood as "that which is most perfect in all of nature" (ST I$^{a}$, q. 29, a. 3). It follows that "possession" of another person, through interpersonal love, be the highest good, and perfect fulfilment of one's being.

According to the hylomorphic and Thomistic perspective, we are rational *because of* our human nature; we are not rational on the basis of performing certain behaviours. TMP proposes that the rational operations of human nature, including the capacity to know reality and make choices, are unique and essential properties of human beings. Therefore a corollary of the notion of participation in being, is that rationality may not be reduced to reasoning processes, nor to subjective self-awareness. Such a focus on behaviours, and quantitative or qualitative processes,

is open to the criticism that it reduces the concept of rationality itself. TMP, on the other hand, maintains that while rationality is manifested in processes which are material, or well as subjective experiences it cannot be reduced to them. (Mullins 2017) Rationality is an essential property inherent to human beings, and TMP offers a most coherent account because it does not reduce rationality to mere processes and experiences.

While subjective experiences are acknowledged, they do not define human existence but rather emerge as consequences of it. Consequently, questions about the hard problem of consciousness hold less significance within the TMP framework, as consciousness is viewed as a consequence of human rational nature rather than a constitutive aspect. (Mullins 2022b)

It follows that any discussion of "rationality" in animals (Edwards and Pratt 2009; Hemingway et al 2017) or in machines, is philosophically questionable. Specific behaviours and operations such as agency, cognitive processing, reasoning, executive control, decision making, intentional goal election, mental intentions, consciousness, qualia, and self-awareness, may only be applied to animals and machines with caveats. It would seem better to use descriptors such as "intelligent".

Certainly by application of the strict test of TMP, rationality can neither evolve, nor be encoded in AI.

## 2. Discussion

This paper finds no philosophical impediments for a Skynet scenario. Such a scenario seems credible in the near future not because machines will have sprung into rationality and self-awareness, but simply because machines can feasibly have reached a complexity in learning and decision making is of a category equal to the appetitive and organisational aptitudes of non-human predatory animals, a far lower threshold for such behaviour than human rationality would provide. To dominate the planet, the snakes need not be rational; nor need Skynet be rational. They learn from sense driven behaviours; Skynet's machine learning is activated by

sensors. The snakes exhibit sense conditioned appetites; Skynet learns through algorithms. A Skynet scenario hypothesises sophisticated behaviours in machines akin to non-rational animal predatory behaviour of the snakes who have, in comparable ways to other species, evolved highly predatory characteristics that, in the absence of natural enemies, dominate their planet.

### 2.1. Common ground

There are similarities between AI and a rationality founded on TMP. Neither dally with dualist explanations. Whatever intelligence is present in AI is wholly and solely materially founded. Within TMP the immaterial operations of rationality can be understood as transcendent actions carried out by ensouled matter through a participated power. They are not structures or representations, nor the result of some ghostly substance or assertions about the existence of inherent properties.

Furthermore AI and TMP hold that there is no mental life without physical structures. But TMP reasons that mental life, while entirely mediated by, is not reducible to, neurobiological structures and processes. This claim is founded on the convertibility of Aristotelian transcendentals: that that participation in truth is founded on participation in being, and is supported by an understanding of the concept of the active intellect as a "participated power." (see 2022b, 180). The contrast with AI is immediately evident. All acts of understanding and reasoning in embodied life are mediated by physical elements, processes, and systems but those elements, processes, and systems cannot fully account for them.

AI and TMP offer different outcomes. Certainly as Aquinas insists, within embodied existence, knowledge is derived from sense data environmental data, but data alone are not knowledge. Sense data are only "the material cause" of intellectual knowledge (ST, I, q. 84, a. 6). TMP presents a comprehensive philosophical basis for analysing insights from neurobiology regarding human behaviour, and for a critique of emergent or code-created physicalist visions of rationality. (cf. Mullins 2022a and b) As such, it stands as a tool to critique claims for machine intelligence.

It offers a more coherent and ultimately a richer explanation for the presence of rationality.

Similarities can be deceptive.

## 2.2. Natural, sensitive appetitive drives move the snakes who dominate the planet

What is the biophysiological explanation for the behaviour of the snakes? First it is to be noted that the behaviour of the snakes derives from sense perceptions that inform appetitive drives in snake brains. These appetitive drives adapt to the environment via learning mechanisms that tailor natural appetites, through neural plasticity consolidated through direct experience as well as epigenetic or evolutionary mechanisms. Aquinas divides such natural appetites according to their objects: according to which an object is perceived as good, and either to be directly obtained or obtainable with difficulty. (ST I, q. 31, a. 5; I, q. 32, a. 5; I–II, q. 23, a. 1; QDdV 25.2). Pleasure or pain are the simple motivations for animal movement.

Physically founded, sensitive appetitive powers are fit for purpose for domination of a physical environment and such appetitive powers are replicable in AI. There is no philosophical impediment to such biological mechanisms being replicated in AI. Sense apprehension would be substituted by sensor registration. Use induced consolidation of neural pathways for simple habit formation, for attentional control, for fear regulation and for conditioned pleasure acquisition mediated by neurotransmitters would be governed by machine learning algorithms. Instinctual and epigenetic evolutionary adjustments also would be direct responses to tangible stimuli in the same ways driving adaptation in other animal species.

Studies of human behaviour draw on the psychological model proposed by Aristotle and perfected by Aquinas. It is reasonable to interpret the snake behaviour as non-rational and operating at the appetitive level. In human beings these appetitive drives have the capacity to be governed in the interests of justice by the development of the cardinal virtues. Such guidance is not available to a non-human animal and will be dependent

in AI on the quality of the programming and the ethical formation of the developers. We will return to this point below.

### 2.3. A participated rationality can neither evolve nor be encoded in AI

Rationality as understood within TMP may never be applied to a machine, on participatory grounds. To recognise the contingency of human life is to acknowledge the reality of participation in being, that existence of the person is imparted in some way from an Ultimate Source. Furthermore the non-material capacity for human beings to grasp in knowledge what is not tangible, and to act in pursuit of non-tangible goals, such as interpersonal love, demand a non-material source. The convertibility of the transcendentals grounded on being, points to human rationality as an essential characteristic of the human person, not simply an arbitrary feature.

On the basis of these complementary approaches I suggest that a Skynet could achieve world domination acting while only in non-rational ways. There is no requirement that rationality be a state of being. Rationality should not be confused with behaviours such as reasoning, tangible goal setting, fear responses, desires, and loyalty to one's species.

### 2.4. Justice guiding all AI development

The Planet of the Snakes thought experiment demonstrates that it is philosophically inaccurate to think of rationality as having some form of threshold, or that "intelligent" behaviours can aggregate to a sufficiently rich level to be considered as a "new order of intelligence", to appropriate words from Terminator character Kyle Reese. Rationality is a state of being not a ticklist of sophisticated behaviours.

On the other hand, rogue AI is conceptually unimpeded by philosophical impossibilities. It is dependent purely on machine development unconstrained by ethical guidance. Such a conclusion leads to reassessed priorities for AI development and ethics.

What ethical formation is required, in complement to developmental strides, to avert a Skynet scenario? Ethical development constraints must

be grounded in an absolute reverence and respect for human life, so that AI development be *unequivocally* good for humanity. For effectiveness, these constraints must be supported by monitoring and legislative safeguards.

There is no consensus in the literature on the form that such safeguards should take. It varies from a grin-and-bear-it approach whereby "citizens can [...] get treated for the potential anxieties that may arise" (Gonzalez-Jimenez 2018,18), to approaches advocating regulation by rules, with some voices also calling for formation of the developers in virtue.

Hinton and Davidson urge regulatory rules for research (Allyn 2023; Davidson 2023). Others seek a multifaceted approach of diversified funding, regulation and dissemination of best practice. (Thais 2024) Božić (2023) suggests that dangers may be mitigated by rules for ethical development, by education and awareness, by transparency, by collaboration between humans and machines, and by research and development.

Francis' 2024 statement voices concerns not only about the dangers of autonomous weapons, but also about the danger that humans themselves settle for an impoverished understanding of what it means to be human.

> Developments such as machine learning or deep learning, raise questions that transcend the realms of technology and engineering, and have to do with the deeper understanding of the meaning of human life, the construction of knowledge, and the capacity of the mind to attain truth. ...not only intelligence but the human heart itself would risk becoming ever more "artificial".

The statement ends with an appeal to pass onto future generations a world of "greater solidarity, justice and peace."

Borenstein and Howard come perhaps closest to this view:

> It is of paramount importance to train future members of the AI community, and other stakeholders as well, to refect on the ways in which AI might impact people's lives and to embrace their responsibilities to enhance its benefits while mitigating its potential harms. This could occur in part through the fuller and more systematic inclusion of AI ethics into the curriculum. (Borenstein and Howard 2021, 61)

Although it is beyond the scope of this paper to map out detail, I suggest that the guiding principle for all AI development must be the virtue of justice, understood in the Thomistic sense, as the virtue of the will, whereby every choice is infused with care and respect for fellow human beings. AI which serves some and harms others offends the common good. More than lip service to Asimov's "Laws of Robotics" is required.

> First Law of Robotics: A robot shall not harm a human or by inaction allow a human to come to harm. (Asimov 2004).

The devil is in the detail. "Harm" includes intangibles. AI must be developed to enhance the relational. It is here that the ethical dilemma of the use of AI resides. The just development of AI requires attention to the common good of all. Common good cannot be interpreted without a worldview encompassing intangible hypotheticals. Common good calculus must be encoded by a programmer who appreciates human fulfilment in truth and love. Anything less than a vision of human persons perfected in truth and in loving relationships between persons, constitutes an impoverished view of rationality. If a programmer lacks this, AI cannot supply what is lacking. Without this rich understanding of human fulfilment, the resultant AI will be a danger. And to lose sight of such human fulfilment, is to lose the grasp of how to get there though the development of virtues, embodied excellences of our human nature. Only by ourselves embodying what rational fulfilment disposes us to, can we defend ourselves against the misuse of intelligence.

## 3. In conclusion

Our excursion to the Planet of the Snakes and reflection on the meaning of rationality serves to remove objections to a Skynet scenario and to bring the timeline forward. It seems reasonable to hypothesise that future machines, utilising sensor data and sophisticated machine learning, but subject to already evident anomalies, such as proxy gaming, goal drift and

optimization of flawed objectives (Hendrycks and Woodside 2003), could indeed operate as predatory animals. Our excursion to the foundations of rationality within TMP argues that there is essentially no metaphysical objection and therefore no philosophical objection to a Skynet scenario.

The response however, I suggest, lies in the human nature itself, perfected within a personalist paradigm of virtue at the service of relationality. To Charles Darwin there was no doubt about the need for an ethics that prioritises "disinterested love", exquisite mutual care and respect for all. We are still a long way from there.

> An anthropomorphous ape, if he could take a dispassionate view of his own case, […] might insist that they were ready to aid their fellow-apes of the same troop in many ways, to risk their lives for them, and to take charge of their orphans; but they would be forced to acknowledge that disinterested love for all living creatures, the most noble attribute of man, was quite beyond their comprehension. (Darwin 1871, 10)

Yes, we must tread very carefully!

## References

Allyn B. 2023. "'The godfather of AI' sounds alarm about potential dangers of AI." Available at: https://www.npr.org/2023/05/28/1178673070/the-godfather-of-ai-sounds-alarm-about-potential-dangers-of-ai (accessed on 06 June 2023).

Aquinas, St. Thomas. *Exposition of the Hebdomads of Boethius* (Introduction and translation by Janice L. Schultz and Edward A. Synan) https://archive.org/details/an-exposition-of-the-hebdomads-by-aquinas_202109/page/n7/mode/2up. EHB in text.

Aquinas, St. Thomas. *Questiones disputatae de veritate*, edited by Joseph Kenny OP. Available at: https://isidore.co/aquinas/QDdeVer.htm. QDdV in text.

Aquinas, St. Thomas. 1955–57. *The Summa Contra Gentiles*, edited by Joseph Kenny OP. New York: Hanover House. Available at: https://isidore.co/aquinas/ContraGentiles.htm. SCG in text.

Aquinas, St. Thomas. 2017. *The Summa Theologica*, translated by Fathers of the English Dominican Province, Online edition. Available at: https://www.newadvent.org/summa/. ST in text.

Aristotle, *Metaphysics*, translation by W.D. Ross. Available at: http://classics.mit.edu/Aristotle/metaphysics.8.viii.html

Asimov, Isaac. 2004. *I, Robot*. New York, Bantam Books,

Bazán, B. C. 1997. "The human soul, form and substance? Thomas Aquinas' critique of eclectic Aristotelianism." *Archives d'histoire doctrinale et littéraire du Moyen Âge*, 95–126.

Božić, V. 2023. *The dangers of artificial intelligence.* General hospital Koprivnica.

Chatila, Raja, Erwan Renaudo, Mihai Andries, Ricardo-Omar Chavez-Garcia, Pierre Luce-Vayrac, Raphael Gottstein, Rachid Alami et al. 2018. "Toward self-aware robots." *Frontiers in Robotics and AI* 5: 88.

Cheney, Dorathy L. 2011. "Extent and Limits of Cooperation in Animals." In *In the Light of Evolution: Volume V: Cooperation and Conflict*, edited by J.E. Strassman, et al. Washington (DC): National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK424864/

Darwin, Charles, 1871. *The Descent of Man*.

Davidson, Tom. 2023. "The Danger of Runaway AI." *Journal of Democracy* 34(4): 132–140. https://dx.doi.org/10.1353/jod.2023.a907694.

De Haan, D. 2018. "The interaction of noetic and psychosomatic operations in a Thomist hylomorphic anthropology." *Scientia et Fides* 6(2): 55–83. DOI: http://dx.doi.org/10.12775/SetF.2018.010.

Dutt, N., Regazzoni, C. S., Rinner, B., & Yao, X. 2020. "Self-awareness for autonomous systems." *Proceedings of the IEEE 108*(7): 971–5.

Fabro, Cornelio. 1966. "The transcendentality of ens-esse and the ground of metaphysics." *International Philosophical Quarterly* 6(3): 389–427.

Fabro, Cornelio. 1969. "Premessa." In *Esegesi Tomistica*. Roma: Libreria Editrice della Pontificia Università Lateranense, Roma, xxxiii.

Fabro, Cornelio. 1970. "Platonism, Neo-Platonism and Thomism: Convergencies and Divergencies." *Neo-Scholasticism* 44: 69–100.

Fabro, C. 1974. "The Intensive hermeneutics of Thomistic philosophy: The notion of participation." *Review of Metaphysics* 27: 451–7.

Feser, Edward. 2005. *Philosophy of mind: A short introduction*. London: Oneworld.

Fjelland, Ragnar. 2020. "Why general artificial intelligence will not be realized." *Humanities and Social Sciences Communications* 7(1): 1–9.

Francis I. 2024. *Message for the World Day of Peace*. https://www.vatican.va/content/francesco/en/messages/peace/documents/20231208-messaggio-57giornatamondiale-pace2024.html.

Future of Life Institute. 22 March 2023. "Pause Giant AI Experiments: An Open Letter." https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Gazzaniga, Michael S. 2011. *Who's in charge? free will and the science of the brain*. New York: HarperCollins.

Gonzalez-Jimenez, H. 2018. "Taking the fiction out of science fiction:(Self-aware) robots and what they mean for society, retailers and marketers." *Futures* 98: 49–56.

Hauser M. 2006. *Moral minds*. New York: Harper Collins.

Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside. 2023. "An overview of catastrophic ai risks." *arXiv preprint arXiv* 2306: 12001.

Hunt, T. 2023. "Here's Why AI May Be Extremely Dangerous—Whether It's Conscious or Not." *Scientific American* 25 May.

Jaworski, William. 2016. *Structure and the metaphysics of mind: How hylomorphism solves the mind-body problem*. Oxford: Oxford University Press,

John Paul II. 1998. *Fides et ratio*. Vatican City: Polyglotta Vaticana.

Koterski, Joseph W. 1992. "The Doctrine of Participation in Thomistic Metaphysics." In *The Future of Thomism*, edited by D. Hudson and D. Moran South Bend. Notre Dame: University of Notre Dame.

Liu, Bai. 2023. "Arguments for the Rise of Artificial Intelligence Art: Does AI Art Have Creativity, Motivation, Self-awareness and Emotion?" *Arte, Individuo y Sociedad* 35(3): 811.

Madden, J. 2013. *Mind, Matter, and Nature: A Thomistic Proposal for the Philosophy of Mind*. The Catholic University of America Press.

Mullins, A. 2016. "Philosophical prerequisites for a discussion of the neurobiology of virtue." *Ethical Perspectives* 23(4): 689–708.

Mullins, A. 2017. "Can neuroscientific studies be of personal value?" *International Philosophical Quarterly* 57(4).

Mullins, A. 2022a. "A Thomistic Metaphysics of Participation Accounts for Embodied Rationality." *International Philosophical Quarterly* 62(1): 83–98.

Mullins, A. 2022b. "Rationality and human fulfilment clarified by a Thomistic metaphysics of participation." *Scientia et Fides* 10(1): 177–95. DOI: https://doi.org/10.12775/SetF.2022.009.

Namestiuk, S. 2023. "Challenging the Boundary between Self-Awareness and Self-Consciousness in AI from the Perspectives of Philosophy." *Futurity Philosophy* 2(4): 43–60.

Norris Clarke, W. 1981. "The natural roots of religious experience." *Religious Studies* 17(4): 511–23.

Norris Clarke, W. 1992. "Person, Being, and St. Thomas." *Communio* 19(4): 616.

Norris Clarke, W. 1995. Explorations in Metaphysics. Southbend: UND Press.

Reuters. 9 June 2023. "Simulation of AI drone killing its human operator was hypothetical, Air Force says." https://www.reuters.com/article/factcheck-ai-drone-kills/fact-check-simulation-of-ai-drone-killing-its-human-operator-was-hypothetical-air-force-says-idUSL1N38023R/

Rziha, J. M. 2009. *Perfecting human actions. St Thomas Aquinas on human participation in eternal law*. Washington: Catholic University of America.

Scharre, P. 2019. "Killer apps: The real dangers of an AI arms race." *Foreign Affairs* 98: 135.

Scruton, Roger. 2014. *The Soul of the World*. Woodstock: Princeton University Press.

te Velde, Rudi. 2015. "Aquinas's Aristotelian science of metaphysics and its revised Platonism." *Nova et Vetera* 13(3): 743–64.

Thais, S. 2024. "Misrepresented Technological Solutions in Imagined Futures: The Origins and Dangers of AI Hype in the Research Community." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 1455–65.

Tredinnick, L., & Laybats, C. 2023. "The dangers of generative artificial intelligence." *Business Information Review* 40(2): 46–48.

Various, 2015. "An open letter, research priorities for robust and beneficial artificial intelligence." Retrieved from https://futureoflife.org/2015/10/27/ai-open-letter/.

Vice: Motherboard blog. 1 June 2023 "AI controlled drone goes rogue, kills human operator in USAF simulated test." https://archive.is/2nE5i#selection-1177.0-1181.311.

Walker, A. J. 2004. "Personal singularity and the communio personarum: A creative development of Thomas Aquinas' doctrine of esse commune." *Communio: International Catholic Review* 31: 457–79.

Wang, J. 2023. "Self-Awareness, a Singularity of AI." *Philosophy* 13(2): 68–77.

Whitehead, Hal. 1997. "Analysing animal social structure." *Animal Behaviour* 53: 1053–67. http://whitelab.biology.dal.ca/hw/Whitehead_Analsoc.pdf.

Wippel, John F. 2003. "Metaphysical foundations for Christian humanism in Thomas Aquinas." *Congresso Tomista Internazionale* (21–25 September 2003), Roma: 4–18.

Wired 28 September 2023. "Six Months Ago Elon Musk Called for a Pause on AI. Instead Development Sped Up." https://www.wired.com/story/fast-forward-elon-musk-letter-pause-ai-development/.