AI and Ethics: Reality or Oxymoron?

Author: Jean Kühn Keyser

Date: 27/07/2023

Preface:

There is a lot of literature to be found regarding the topic of AI and Ethics, much relating to the potential and threats the latter poses to our legal, socio economic, environmental, financial, and political life[1], the difficulties of which has been highlighted recently by just how laborious the process of passing new laws about AI and Ethics has been for the EU. Much of what ailed the process had to do with merely defining what AI is and should constitute, as different groups struggled to delineate between essentialism viruses a more expanded definition[2]. However, when one looks more closely at the debate you will notice what is missing is AI itself, which will then be our point of departure.

Before we start I feel we need to make it clear that for the sake of this piece we will not refer to the idea of AI in terms of sentience[3] as this relates to a state of metaphysics which has historically not only been used to create classes of being (and justified cruelty towards animals for example), but it did so based on a state of existence which we have not been able to reasonably prove outside the realms of religion. Said differently, if we consider Wittgenstein, metaphysics, epistemology, and ontology each play their own language game which it irreconcilable[4]. It is also worth noting that this article will read much like a literature review, as is often the case in philosophical investigations into already exiting text.

Introduction:

In Hagendorf's article he critiques the field of AI ethics for not living up to its own standards, the author concludes by stating that AI ethics poses several risks to itself ranging from "*risks that* […] *originate from ethicists themselves or from the consequences their embedding in AI organizations has*" (Hagendorff, 2022: 6). They attribute these risks to an array of reasons relating to the human element of AI ethics and practitioner themselves, from the "*psychological considerations about bounded ethicality in ethicists themselves,* […] *considerations regarding the individuals who react to (or ignore) ethical principles and advice, to the complicated professional role of*

[1] This is especially true with regards to what is called tort law which applies when *"someone commits a wrong against another person. Tort law allows individuals who have had a wrong committed against them to claim damages against the person who has committed the wrong*" (Available online: https://www.thelawyerportal.com). It encompasses a wide variety of different types of legal issue is part of our civil law that aims to return individuals back in the position they were in before the wrong was committed. Hence tort law includes: "[d]*uty of care*" where the hard was reasonably forceable and there are some relations between the parties; "[n]*egligence*", when a person fails in their duties of care; "[p]*ersonal injury*" which are brought to order to establish compensation for injuries sustained; "[s]*trict liability*" which exist usually where a behaviour is inherently dangerous, in which case "*the individual claiming damages only needs to prove that the tort occurred, and the defendant was responsible*"; and lastly "[n]*uisance*" which can be private or public. Private nuisance refers to "*situations where actions by the defendant causes unreasonable interference with a private individual's land or enjoyment*" thereof. Public nuisance similar "*except the action of the defendant interferes with a group rather than just an individual*" (Available online: https://www.thelawyerportal.com).

[2] We will look into this topic a bit more in Section two.

[3] Trasnshumansim is "*the belief that technology can transcend the limitations of the human body and brain*", according to which AI is a key component (Hughes, 2009: 1). Ironically then even though "*most transhumanists are atheist their belief in the transcendent power of intelligence generates new theologies"* (2009: 1)

[4] This is in accordance to Wittgenstein's language philosophy where he propounded that different forms of knowledge platy different types of language games and so it makes little sense to try and debate science with religion or vice versa (1922).

AI ethicists" and the types of AI ethics policies, audits etc. they deal with (2022: 6).

Having said this, in this piece we will considering these issue along linguistic philosophical lines:

- Considering the problems of defining delineations for AI and AI regulation (which will be discussed in section two with regards to the EU) and;
- Problamatize the use of the term AI ethics itself, pointing to the fact that from the outset that what is missing is 'AI' itself.

Methodology:

To better delve into this question methodologically will require critical self-reflection[5], applying negative dialectics to deliberate if such an idea such as AI even exists in praxis. This would be most appropriate in delving into: if part of what ails this field of study is exactly the insistence of ascribing latter terminology to what actually amounts to a broader field of study, and considering if AI ethics as it is ascribed to currently, is congruent with the emergence of such a consciousness or 'being'. The reason for using negative dialectics will hopefully become evident as we go forward, but for now the broader idea relates to the contrast of the latter with what Adrono calls positive identity.

In brief, positive identity signifies the process whereby "*human thought, in achieving identity and unity, has imposed these upon objects, suppressing or ignoring their differences and diversity*", in this case the instance of usage AI in the absence of sufficient proof of existence (Zuidervaart, 2015: Available online: http://plato.stanford.edu). Adorno believes that we do such things because of people's "*societal formation whose exchange principle demands the equivalence (exchange value) of what is inherently nonequivalent (use value)*" (2015. Available online: http://plato.stanford.edu). What this means is that there is "*the "application*" of *a priori* concepts to priori intuitions via the "*schematism" of the imagination (Einbildungskraf)*" (Zuidervaart, 2015. Available online: http://plato.stanford.edu). Adorno calls this "*constitutive subjectivity*" (Adorno, 1973: xx). This kind of identity giving assumes that a 'thing's' identity is the "*thing in itself*", or simply by applying the term AI to functional algorithm enough times, it somehow exists (2015. Available online: http://plato.stanford.edu). It is the *reification* of AI[6].

Subsequently, to see if this is the case we will look at historical narratives concerning AI, which will potentially prove firmer grounds for analysis. Aligned with mentioned methodology, starting from within, we will consider why is it so difficult to delineate boundaries of consciousness when talking about AI, driving one faction to metaphysical sentience[7] and those more akin to policy writing to what appears to be pure refutation?

As will become evident, this split in opinion has to do ontologically, with how AI appears to us as a product of mathematics, and how as a "*mathematical object* [it] *is neither transcendent nor immanent*" (Badoiu, 2006: 45). Said differently, mathematical objects are pseudo beings "*suspended between a pure separate act, whose supreme name is God, and*

[5] Critical self-reflection approaches ethics "*not merely* […] *from the external perspective of applied ethics – in favour of ethical correctives against economic rationality, but instead does it – from the internal perspective of a self-reflective way of thinking*" (Zoalnai, 2004: 17). As a method of analysis in this context it means looking at the internal contradictions of thought regarding the use of the concept AI, which necessitates "'*determinate negations' pointing* [out] *specific contradictions between what thought claims and what it actually delivers*" (Adorno qtd, in Zuidervaart, 2015. Available online: http://plato.stanford.edu)

[6] Reification is the "*act of changing something abstract* ([…] *existing as a thought or idea) into something real*" (Available online: https://dictionary.cambridge.org)

[7] As described in footnote 3.

sensible substance, or actually exiting things [… it] *is neither physics nor metaphysic*" (2006: 45). AI appeals to us exactly because it speaks to our own vanity, it is based on learning algorithms that seem to mimic our own cognitive processes. Ironically however, this vanity is at odds with itself given according to the historical definition of AI, it needs to be removed far enough from our own human intervention as to make the latter inconsequential, only once this is achieved can AI claim to be a separate 'new kind of being', but more on this later. In evolutionary terms one may say 'its needs only be different enough to guarantee that it is something separate from where it came from', like the small yet crucial genetic differences between humans and Bonobos.

Structure of this article:

Given the broad scope of what is being studied, this article will be divided into three parts: Firstly, we will look at the historical emergence of AI as a concept and its delimitations as existing as a potential moral agent.

Secondly, we will consider the issues of allocation related to the use of the term AI as it is currently implemented in many ethics policies.

Thirdly and lastly, we will contemplate what the possibility is of AI in the future as a potential moral agent and subsequent implications this may hold in terms of ethics and ethical policies.

Section One: Historical overview

Starting from a contemporary place in time, the reason why this topic has again come to the fore again has to do with the developments in Large Language Models (LMM's), and specifically ChatGPT where claims reading the emergence of AI have yet again surfaced. However, buy taking a step back it is hoped we can delineate how much of the latter has any real validity, and how much of AI is still merely "*rigorous aesthetics*", lacking the experiences of a "*real-being*" required for understanding the consequences of taking personal responsibility for ethical choices (Badoiu 2006: 48).

To begin with we have to look at Allen Turing who delineated AI at minimum as something which possesses the properties of mere emergent consciousness. From there we will look at later distinctions which scrutinized AI in relation to ethics, some insisting on it having a physical presence if it was to assume the moral agency[8] prerequisite for ethical responsibility.

Turing's test for detect the existence of AI is basically:

> Could a machine convince a human judge, who poses questions to it behind a closed compartment through text alone (so as is not to get any visual cues apropos who or what they are questioning), after repeated 5 minutes of questioning, 70% of the time that they are not certain if it is a machine or not to whom they are speaking.

[8] Moral agency refers to the individual who has the *"capability for 'creativity', regarding discretionary judgement and taking responsibility for moral decisions as determinant of their moral agency*" (Keyser, J. 2009: 22). Even though the focus is on moral agency as it precedes legal identity, what this means in terms of law is competence. In forensic psychiatry this competence is defined as "*a designation used of a person who has been judged mentally capable to stand trial*" (Rebber, 1985: 137). The criteria being "*(a) the person understands the nature of the charge and the legal consequences of adjudged guilt; and (b) the person is able to assist in his or her defence*" (1985: 137).

Put more simply: "[w]*e place something behind a curtain and it* [communicates] *with us. If we can't make the difference between it and a human being then it will be A*I", this including parts of human speech that may be considered more general or less intelligent (Wagoner, 2004: 4).

It may seem simple enough, particularly since from within this line of reasons all that would be required from an AI to be considered ethically responsible, would be the ability to converse with humans on an equal footing. A sort of disembodied consciousness which needs no real physical presence. However, even this is not as easy as one would assume, as will be discussed at the end of this section: the process of conversing with another human intelligibly requires the creative[9] ability to use both context specific and universal understandings to form meaningful dialogue.

Given that the original Turing test did not speak directly to ethical and moral concerns the "*comparative moral Turing test*" (*cMTT)"* was devised (Anderson and Anderson, 2007: 24). This to some extent did away with the problem relating to "*disagreement concerning definitions of ethical behavior as well as the requirement that a machine have the ability to articulate its decisions*" (2007:24). In this revised format of the Turing test:

> *"an evaluator assesses the comparative morality of pairs of descriptions of morally significant behavior where one describes the actions of a human being in an ethical dilemma and the other the actions of a machine faced with the same dilemma. If the machine is not identified as the less moral member of the pair significantly more often than the human, then it has passed the test*" (Anderson and Anderson, 2007: 24)

It was argued that his would be a sufficient experiment given humans' ability to make ethical decisions is not perfect. However, the hurdle of applying common sense and recognition of fallibility versus mere miss calculation would persist. Moreover, if we are talking about the use of AI in situations where people's lives may be at stake and *tort* law comes into play, "*a machine that passed the cMTT might still fall far below the high ethical standards to which we would probably desire a machine to be held*" (Anderson and Anderson, 2007: 25).

I the face of this some have argued that when it comes to AI and ethics, a disembodied intelligence is not enough since just as "[h]*uman intelligence is diversified into human activity*" and within this specific works context, so too "[m]*oral and political intelligence are also predominantly action-oriented*" (Pana, 2006: 257). Hence in terms of an AI's ability to hold moral agency, it would have to prove to be capable of physically autonomous action (Pana, 2006: 257), some of which we already observe in the form of for example "*autonomous drones*" (Dawes, 2022. Available online: https://robohub.org) and automated financial investing.

In congruence with this line of reasoning (given that human lives and financial security could be at risk), it is argued that for an AI to truly mimic the human cognition and ethics

[9] Creativity be defined as the "'*mental processes that lead to solutions, ideas, conceptualizations, artistic forms, theories or products that are unique and novel'"* (Rebber, 1985: 165), it necessitate higher cognitive processes such as imagination, "*which in turn is only possible given the existence of"* some sort of individual with lived experiences (1985: 345). This is because *"the imagination is the process of recombining memories of past experience and previously formed images into novel constructions*" (1985: 345).

requisite to hold it responsible as a moral agent, nine additional criteria are necessary. The mains reasons for these additional criteria comes down to the ethical issues pertaining to ethics of rights[10] as well as justice[11]. Specifically: retributive and distributive justice in relations to holding a moral agent personally responsible for ethical conduct; negative rights when it comes to not inferring with a moral agent's ability to make choices and subsequent consequences that may follow; and lastly positive rights with respect to who is responsible for providing additional rights where the moral agent is sufficiently different to warrant it for the sake of social equity. As sch these additional criteria for something to be considered an AI it would have to prove that it is:

> 1 – an individual entity with "*complex, specialized, autonomous or self–determined, even unpredictable conduct"* (Pana, 2006: 254).
>
> 2 – an entity which is *"endowed with diverse or even multiple intelligence forms, like moral intelligence"* (2006: 254).
>
> 3 – able to act as an *"open and, even, free-conduct performing* [system] *(with specific, flexible and heuristic mechanisms and procedures of decision)"* (2006: 254).
>
> 4 – a system open to learning and being educated, not just merely following instructions (2006: 254).
>
> 5 – a system that has a *"lifegraphy", not just "stategraphy"* (2006: 254).
>
> 6 – holds beliefs and not mere automatisms (2006: 254).
>
> 7 – since a moral life has a form of spiritual and not merely conscious activity, it should be *"capable even of reflection"* (2006: 254).
>
> 8 – be part of or have *"elements/members of some real*

[10] Ethics of rights in most basic terms can be understood as "*a justified claim against another person's behaviour - such as my right to not be harmed by you. Rights and duties are related in such a way that the right of one person implies the duties of another person*" (Fieser, 2006: online). These duties can then be divided into positive and negative rights. Positive rights are "[d]*uties of other agents (it is not always clear who) to provide the holder of the rights with whatever he or she needs to freely pursue his or her interest* [therefore] *do more than impose negative duties. They also imply that some other agent* [or institution], *have the positive duty of providing the holders of the right with whatever they need to freely pursue their positive rights*" (Velasquez, 2006: 76). Conversely, negative rights are what we normally attribute deontological ethics with and are defined as "[d]*uties others have not to interfere in certain activities of the person who holds the right distinguishing the fact that its members can be defined wholly in terms of the duties others have not to interfere in certain activities of the other person who holds a given rights*" (2007: 76).

[11] Ethics of justice can be divided into three parts: "*distributive justice which concerns how we should distribute the products of social cooperation among the community's citizens*" (LaFollette, 2002: 511". More succinctly the latter refers to the "[d]*istributing of society's benefits and burdens*", the fundamental principal being that "[i]ndividuals who are similar in all respects relevant to the kind of treatments in question should be given similar benefits and burdens even if they are dissimilar in other irrelevant respects; and individuals who are dissimilar in a relevant respect ought to be treated dissimilar, in proportion to their dissimilarity" (Velasquez, 2007: 88-89). In addition to this we have "*retributive justice* [which relates to] *blaming or punishing persons fairly for doing wrong and* [lastly] *compensatory justice* [that refers to] [r]*estoring to a person what the person lost when he or she was wronged by someone*" (Velasquez, 2006: 88). Ethics of justice and ethics of rights are inexorably linked, however ethics of justice generally do not "*override the moral rights of individuals,* [because] *to some extent, justice is based on individual moral rights*" (2006: 88). The reason being that "*the moral rights of some individuals cannot be sacrificed merely to secure a somewhat better distribution of benefits for other*" (2006: 88).

(corporal or virtual) community" (2006: 254),

9 – as cultural beings have *"free conduct* [which] *gives cultural value to the action of a" natural" or artificial being"* (2006: 254).

We find then here a crucial delineation in terms of AI, moral agency, and the ability to make ethical decisions. In simple terms is comes down to how AI should defined be as an "*explicit ethical agent*"[12] (Anderson and Anderson, 2007: 15), which not only solves "*difficult problems for the society but also reproduce mentality in machines"* (Nat and Sahu, 2020: 105)

The revised Turing test would fall under what is called the functionalism perspective where for an AI to have moral agency all that is required from an AI is "*the presence of certain behaviours and reactions that are functionally equivalent* (Wallach and Allan, 2008) *to the behaviours and reactions which advocates of the standard view would view as mere indicators of standard criteria*" as broadly set out above (Behdadi and Munthe, 2020: 197).

Equally, the second group which holds the *standard view*, has refined the initial nine criteria down to "*rationality, free will, and phenomenological consciousness*" (2020: 197).

According to the standard view an entity with moral agency should be able to:

1. Cause physical events with its body (Behdadi and Munthe, 2020: 198).
2. Have an "*internal state, I, consisting of its own desires, beliefs, and other intentional states that together comprise a reason to act in a certain way (rationality and consciousness*)" (2020: 198).
3. The state of I is "*the direct cause of 1*" (2020: 198).
4. Events in 1 "*has some effect of moral importance*" (2020:198).

So, it is no longer enough for an AI to be disembodied, but it should have some recourse to physical actions based on internal states that have real moral consequences. This is important because it is argued that ethical and moral actions are not merely causal, but explained and judged in terms of "*our internal mental states*" (Johnson 2006: 198 qtd in Behdadi and Munthe, 2020: 198). To bring back the human element, in terms of the law it refers to the competence of a moral agent.

A counter point made by functionalism proponents are that such standards for AI moral agency is very low (akin to that of an adult human), where as a "*mind-less morality*" could raise this level and make for a less anthropocentric perspective whilst "*maintaining consistency and relevant similarity concerning the underlying structural features of paradigmatic human moral agents*" (2020: 198). They give a separate set of criteria for moral agency:

1. Interactivity: A moral entity should be able to interact with its environment (2020: 199).

[12] There is then a difference between a machine that acts with implicit versus explicit ethics which. Moore makes the delineation that "*a machine that is an implicit ethical agent is one that has been programmed to behave ethically, or at least avoid unethical behavior, without an explicit representation of ethical principles. It is constrained in its behavior by its designer who is following ethical principles. A machine that is an explicit ethical agent, on the other hand, is able to calculate the best action in ethical dilemmas using ethical principles. It can "represent ethics explicitly and then operate effectively on the basis of this knowledge*" (Anderson and Anderson, 2007: 15)

2. Independence: Such an entity should have the ability to "*change itself and its interactions independently of immediate external influence*" (2020: 199).

3. Adaptability: A moral entity may "*change the way in which 2 is actualized based on the outcome of 1*" (2020: 198)

Above and beyond the fact that ethics requires some level of common sense, which prerequisite imagination to develop a common understanding just to have a mere conversation, there an added layer of creativity given that "[u]*nlike many other knowledge bodies, ethics in intricately* [is] *connected to our ability to have firstperson perspectives",* so we may have insight as to how our decision affects others (Nath and Sahu, 2020: 106). In fact Levinas argued that ethics only exists in the presence of *the other* (Thomas, 2004: 87)

Furthermore, from a technical perspective the added cognitive processing that such creativity implies means taking into account additional consideration which are: *Explainability* and *interpretability* in relation to black and white box systems. Though these two terms are often used by researcher interchangeably with no mathematical definition for either one, nor any matrix of measure to distinguish or clarify them and related terms such as comprehensibility, when it comes to learning algorithms they are understood in distinctly separate ways (Linardatos; Papastefanopoulos and Kotsiantis, 2021: 2-3).

Interpretability is connected "*with the intuition behind the outputs of a model* […]*, the idea being that the more interpretable a machine learning system is, the easier it is to identify cause-and-effect"* associations of a system's inputs and outputs (2021: 2-3). On the other hand, explainability is connected with *"the internal logic and mechanics that are inside a machine learning system"* (2021: 2-3). Subsequently*, "*[t]*he more explainable a model, the deeper the understanding that humans achieve"* with regards to knowing the *"internal procedures that take place while the model is training or making decisions"* (2021: 2-3).

This distinction is crucial since "[a]*n interpretable model does not necessarily translate"* into one where humans can understand *"the internal logic of or its underlying processes"* (2021:2-3). Subsequently, for machine learning and learning algorithms "*interpretability does not axiomatically entail explainability, or vice versa" (2021: 2-3).* Therefore, we must say *"that interpretability alone is insufficient and* […] *the presence of explainability"* is of essential importance, as the aforementioned should be consider a broader term than the latter (2021: 2-3).

A consequence of this is that there is a "*trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions"* (2021: 1). Because of this 'trade-off' it means that often a surge in performance is "a*chieved through increased model complexity, turning such systems into"* what is called *"black box" approaches"* (2021: 2-3). They are attributed this title since the produce results that are not easily explainable due to the complexity of their internal workings and potentially having to calculate for variable which even the programmers had not fully taken cognizance. In turn this may cause "*uncertainty regarding the way they operate and, ultimately, the way that they come to decisions"* (2021: 1).

In contrast to this there are also "*white box systems or glass-box models, which easily produce explainable results",* usually including common examples such as *"linear [11] and decisiontree based [12] models"* (2021: 1-2). However, such models are innately less powerful and *"fail* [to] *achieve state-of-the-art performance* […] *compared"* to *black box systems* because of their *"frugal design" (2021:1-2).*

One outcome from this is that those "systems *that cannot be well-interpreted",* but produce high performance are non the less difficult to trust *"in sectors, such as healthcare or self-driving cars, where also moral and fairness issues have naturally arisen"* (2021: 1-2). There is then a conflict in creating high performance systems that are also robust, trustworthy and fair enough "*for real-world applications",* leading to a *"revival of the field of explainable Artificial Intelligence […] focused on the understanding and interpretation of the behaviour of AI systems"* (2021: 1-2).

The biggest concern then for ethicists and policy writers relates to how do we reconcile this internal conflict posed by *black box systems*. Or said differently, concerning AI and deep learning models, how do we produce such systems so they may be trustworthy, robust and reliable without losing performance? Additionally, when considering the parameters set by both the *standard* as well *functionalist's* view as to what the prerequisites for an AI would have to be, how can we achieve any of them if the complexity and performance required to do so intrinsically leads to greater uncertainty and distrust?

There are two additional arguments around the demarcations for the emergence of AI as a moral agent which could make ethical decisions, that relates to the human element to consider. Firstly, the epistemological pragmatic argument which states that if a machine can act "*as-if*" it has moral agency, then it should be considered and moral agent (this seems to be an approach also held by many of the institutions mentioned) Behdadi and Munthe, 2020: 199). Secondly, the arguments related to the independence criteria which both views hold. The problem is the fact that AI is designed by a human and as such:

> "[n]*o matter how independently, automatically, and interactively computer systems of the future behave, they will be the products (directly or indirectly) of human behavior, human social institutions, and human decision*" (Johnson, 2006: 201 qtd in Behdadi and Munthe, 2020: 200).

The consequence of this being that for a true AI to be an *explicit moral agent* it would have had to be created along a very long line of generations of self-generating programs without any human interventions to such an extent that the latter's insets are trivial.

This would constitutionally be a different kind of consciousness with its own unique set of behaviors, social institutions and decision-making abilities and processes. In contrast to this, some have propounded the idea that even though it would be difficult to argue that one would be able to hold machines as moral agents responsible in terms of free will and intentionality "*neither attributes is necessary to do the morally correct action in an ethical dilemma and justify it*", stating it would only have to justify its actions citing acceptable ethical principles (Behdadi and Munthe, 2020: 204). However, when it comes to developing policies and laws around AI, then according to ethics of justice: if and AI is not an *explicit agent* the problem of accountability would always fall on the human programmer who as

such holds *special office*[13] cannot claim ignorance or inability to act. Additionally, the process of ethical deliberation is being overly simplified, as ethical dilemmas are exactly the types of situations in which '*what constitutes the acceptable principles*' are not clear.

Lastly, bringing the historical context, to bear on our contemporarily situation where a lot has been said about recent claims around LLM's and ChatGPT etc. as a potential Ghost in the machine-like emergent AI[14], it has been shown through experimentation that they don't "*learn to emulate reasoning functions. Instead, they find clever ways to learn statistical features that inherently exist in the reasoning problems*" (Dickson, B. 2022. Available online: https://bdtechtalks.com). More succinctly, these models have "*learned to use statistical features in logical reasoning problems to make predictions rather than to emulate the correct reasoning function"* (2022. Available online: https://bdtechtalks.com)

What this means is that when we compare such models with the commonsense nature of actual human communication (as suggested in the Turing test), requiring common sence reasoning rather than merely immolating statistical features, they do not fare well.

When we are talking about commonsense here, what we are referring to within the confines of a conversation is the ability to use logic in a reflective creative way to 'reason'. It is through this reflexive creativity that we can imagine what '*the others'* perspective may be and by combining this with our own universal sets of social knowledge (which we gain through experience inactive or vicariously), we contextualize our conversations and give responses which we hope is interpretable and understandable by the receiver. Through this combination of logic and creative imagining of universal and common experience we create a 'common understanding'.

However, through even humans fail many times in this, we at least have the faculties to move beyond mere immolations of statistical sets which allows us to often to "*omit* […] *shared knowledge*" (2022. Available online: https://bdtechtalks.com). The closest analogue one could imagine in computing is the idea of *superpositioning* in *Quantum bits (Qbits)*[15], and just as with the latter, we often

[13] In ethics 'special offices' refer to the fact that those who hold positions of power or specific responsibilities (such as managers or programmers), cannot rely on the usual arguments of ignorance or inability to act because of lack of power, arguments for mitigating circumstances relating to their actions, given their very position means that they ought to have known and therefore would be remiss in their responsibilities if they were ignorant of did not act with regards to something they have responsibility towards.

[14] The idea that AI will emerge from something like the internet of Google itself is not a new one. The concept being that search engines function like our subconscious "*continuously crawling the internet, indexing every bit of information it can* […] trying *to make sense of this information, attempting to understand its meaning and how it relates to the search queries its users are inputting billions of times a day*" (Adams, 2010. Available online: https://www.stateofdigital.com). Like our subconscious these engines are vastly powerful "*yet unaware of its own existence*" (2010: Online). This is what is often called the Ghost in the machine hypothesis where "*an intelligence-gathering and manipulation program that* [becomes] *self-aware"* (2010: Online). In line with this angle of reasoning, if we combine such a search engine with the computational powers of Qbits, according to this conception of AI we humans become subverted to merely become its senses.

[15] A qubit (or quantum bit) is "*the quantum mechanical analogue of a classical bit. In classical computing the information is encoded in bits, where each bit can have the value zero or one. In quantum computing the information is encoded in qubits. A qubit is a two-level quantum system where the two basis qubit states are usually written as \left\lvert 0 \right\rangle/0) and \left\lvert 1 \right\rangle/1). A qubit can be in state \left\lvert 0*

don't know if a common understanding between parties have been reached, if we don't check if the value or key concepts of the conversation was mutually understood by both parties.

According to Dirksen, it is exactly this ability to move from specific to common sense universal sets of data on which we rely that such *LLM's* have the most difficulty with and fail. Their experiments showed, even though "*machine learning model managed to achieve near-perfect accuracy on one data distribution*", the latter does not "*generalize to other distributions within the same problem space",* even when the *"training dataset"* covered the *"entire problem space and all distributions being derived from the same reasoning function*" (2022. Available online: https://bdtechtalks.com).

Some may argue that by merely enlarging the model we could do away with this problem, but in fact what experiments have found is that "*the logical reasoning problem does not go away as language models become larger* [it..] *just becomes hidden*", as in our *black box* scenario (2022. Available online: https://bdtechtalks.com). So even though "*LLMs can spit out facts and nicely stitched-together sentences,* […] *when it comes to logical reasoning, they are still using statistical features to make inferences, which is not a solid foundation* (2022. Available online: https://bdtechtalks.com). The problem then of reflexivity and *superpositioning* remains and so on the one hand "*when a model is trained to learn a task from data, it always tends to learn statistical patterns"* which already inherently exists in the examples, however *"on the other hand,* […] *the rules of logic never rely on statistical patterns to conduct reasoning*" (2022. Available online: https://bdtechtalks.com). Per se, since it would be almost imposable to construct "*logical reasoning dataset that contains no statistical feature*, *it follows that learning to reason from data* [alone] *is difficult*" without the ability for creative reflexivity (2022. Available online: https://bdtechtalks.com). Furthermore, based on Gödel's incompleteness theorem, we can categorically state that without the *Qbit's* ability for *superpositioning*, normal formal algorithms will never be able to perform tasks of reflexivity (Hofstadter, 1999).

Suffice to say that regardless of if one holds the *functionalist* or *standard* view on AI and moral agency, in both cases AI has *explicit agency*. Having said this, when considering the problems posed by the *superpositioning* and *reflexive creativity* needed for common sense reasoning, we have yet to pass even the Turing test. Additionally, issues relating to the internal contradictions throw up by the complexity vs. performance needs posed by the necessity for *explainable* and *understandable* learning system models (to prevent or lessen *black box system*s), seems to suggest that its very unlikely we have achieved the additional criteria required for an '*explicit agent*' that could be socially trusted, at least no within our human community. In fact, "[e]*nsuring that a machine with an ethical component can function autonomously in the world remains a challenge to researchers*" (Anderson and Anderson, 2007: 25)

To conclude this section, in a recent interview held with Michel Jordan, a leading researcher at Berkley, he stated that in terms of AI they are "*nowhere near advanced enough to replace humans in many tasks involving*

\right\rangle/0⟩, \left\lvert 1 \right\rangle/1⟩ or (unlike a classical bit) in a linear combination of both states. The name of this phenomenon is superposition" (Available online: https://www.quantum-inspire.com).

reasoning, real-world knowledge, and social interaction", and even though some of these systems may show "*human-level competence in low-level pattern recognition skills*" they are merely "*imitating human intelligence, not engaging deeply and creatively*" (Pretz, 2021: Available online: https://spectrum-ieee-org.).

He continues to say that the mimicking of human thought (as we have seen from the historical narrative), is not even the most important goal for machine learning engineers and that "[p]*eople are getting confused about the meaning of AI in discussions of technology trends*" (2021: Available online: https://spectrum-ieee-org).

Section two: The use of the terms AI in ethics policies

In the absence of the actual exitance of AI we need to deliberate:

- What is being said about AI in AI ethical policies?
- Are these policies in correspondence with our given knowledge of what constitutes an AI as something with *express agency*, versus the mere machine learning of Robots, as this has a direct impact as to attribution of burdens and retribution corresponding to *tort* law?

In what is to follow we will look at UNESCO's AI Ethics policy as the example of this double speak for two reasons: Firstly, as a key Ethics policy it will be replicated in many industries and parts of the business sector. Secondly, because it will not be physically possible to look at all possible examples, though there are many. Our point of departure will be UNESCO's[16] anthropocentric conception of AI.

UNESCO's notion of AI ethics is captured in their statement that their "*vision* [is] *for a human-centered AI future,* [as such] *a normative instrument on the ethics of AI should serve as a means of mainstreaming universal values into AI systems, which must be compatible with internationally agreed human rights*" (Garcia, 2021. Available online: https://thegoodai.co).

UNOSECO definition of AI states that it refers to:

> "*systems* [of technology] *which have the capacity to process data and information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control"* (Anon. 2021. Available online: https://unesdoc.unesco.org).

They denote three key elements to their approach

> *":[…] AI systems are information-processing technologies that embody integrate* […] *models and algorithms that produce a capacity to learn and to perform cognitive tasks leading to outcomes such as prediction and decision-making in material and virtual environments"* (Available online: https://unesdoc.unesco.org).

[16] Other similar example can be sought from sources such as "*AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations*. (Floridi; Cowls; Beltrametti; Chatila; Chazerand; Dignum; Luetge; Madelin; Pagallo; Rossi; Schafer; Valcke and Vayena, 2018).

> Such systems "*are designed to operate with varying degrees of autonomy by means of knowledge modelling and representation and by exploiting data and calculating correlations"* (Available online: https://unesdoc.unesco.org). This may "*include several methods, such as but not limited to:* […] *machine learning, including deep learning and reinforcement learning, and(ii) machine reasoning, including planning, scheduling, knowledge representation and reasoning, search, and optimization"* (Anon. 2021. Available online: https://unesdoc.unesco.org)
>
> Lastly, *according to UNESCO "AI systems can be used in cyber-physical systems* [such as] *the Internet-of-Things, robotic systems, social robotics, and human-computer interfaces* [including] *control, perception,* [and] *the processing of data collected by sensors",* furthermore physically is can also mean the *"operation of actuators in the environment in which AI systems work*" (Available online: https://unesdoc.unesco.org)

They define their approach to AI ethics as systematically normatively reflective:

> *"based on a holistic, comprehensive, multicultural and evolving framework of interdependent values, principles and actions that can guide societies in dealing responsibly with the known and unknown impacts of AI technologies on human beings, societies, and the environment and ecosystems, and offers them a basis to accept or reject AI technologies*" (Available online: https://unesdoc.unesco.org).

Clearly sticking with the deontological utilitarian[17] approach which historically has dominated AI ethics, but more on this later.

Having given the main components of UNESCO's views in AI, it seem that on the face of it they seem fair, but realistically this could only be applied to a machine with *implicit agency*. In other words Robots like machines or algorithms. In addition to this, the ethical standards it sets for AI (when read against the background of the revised Turing test), is far higher than it would be for any human for whom such high criteria would be an ideal to be achieved, and which historically we have failed to accomplish many times.

They also completely dispel with the idea that an AI as an autonomous conciseness would ever exist through statements saying that "*humans are ethically and legally responsible for all stages in the life cycle of AI systems* [subsequently], *it follows that an AI system "can never replace ultimate human responsibility and accountability*" (Available online: https://unesdoc.unesco.org). They go further on to say that "*as a rule, life and death decisions should not be ceded to AI systems*" Available online: https://unesdoc.unesco.org), yet (as will be discussed) here already in terms of the use of killer drone this is not happening, and countries using them do not wish to agree to this limitation (Dawes, 2022. Available online: https://robohub.org).

[17] See page 13 with reference to Pereira and Saptawijaya.

From this we can also see that there is a strange kind of double speak going on in that they attribute '*express AI*' characteristics such as *reasoning* which goes further than what we know it foreseeably possess, but then limit AI's autonomy to that of merely an *implicit agent* which is historically not what an AI would be.

Applying the chosen methodology, we can phrase it as:

When we look at the narrative of AI compared to the ethics and expressly policies related to it, there is an internal contradiction. On the one hand they are attributing *a priory positive identity* to current "*elegant algorithms*" (Lakoff, 2009: 3) as if they are talking about the actual existence or emergence of *express AI* properties such as *reasoning* and *perception*, but then seem very "*speciesists*" with a clear "*prejudice or attitude of bias toward the interests of members of* [our own] *species and against those members of other*", i.e. AI (Singer 1975 qtd. in Anderson and Anderson, 2011: 288).

From the outset their anthropocentric point of departure is made clear through stating they are promoting a *human-centered approach* to AI (Available online: https://digital-strategy.ec.europa.eu). They are then trying to play the game both ways, bending the historical narrative of AI as *explicit moral agent*, by using terms such as "*functional AI*" to circumvent the quandary of problems that still need to be resolved to even talk about the emergence of AI. According to this approach by UNESCO, there will never truly be an 'AI' as it's clear their conception of the latter only ever refers to *implicit* learning algorithms that do not stray too far into the realm of *black box systems*.

In linguistic philosophical terms, they are misappropriating the term, as Prof Jordan noted when speaking of machines that could imitate human cognition: "*We don't have that, but people are talking as if we do*", or as with the examples given, attribute qualities related to implicit ethical machines to AI (Pretz, 2021: Available online: https://spectrum-ieee-org.cdn.ampproject.org).

It could be said that, though these institution are considering how "[m]*achine learning* […] *can provide new services to humans in domains such as health care, commerce, and transportation, by bringing together information found in multiple data sets, finding patterns, and proposing new courses of action*", which in and of itself is a noble and worthwhile endeavor, the term we are applying should not be AI as it conflates matters that should not be put together lightly (Available online: https://spectrum-ieee-org.cdn.ampproject.org). This is supported by by Garry Marcus in their article of June 6 2022 stating: "*We are still stuck on precisely the same challenges that academic scientists (including myself) having been pointing out for years: getting AI to be reliable and getting it to cope with unusual circumstances*" (Marcus. 2022. Available online: https://www-scientificamerican-com)

From the human perspective, UNESCO also makes the very idealistic implicit claims that one could develop a singular universal idea of ethics, and that this will somehow make us able to deal with the unknown consequences of AI. Once again bringing us back to the initial problem statement: part of why there has been such a difficulty with delineating AI with regard to laws, is because we are not speaking of AI, but rather we trying to assign in a very anthropocentric manner *a priory positive identity* to very sophisticated *elegant algorithms*.

In term of the law the reality is that technology is developing faster than what it and policies about AI is able to keep up with, simply stated "[e]*thics and jurisprudence, and hence legislation, are* […] *lagging much behind in adumbrating the new ethical issues arising from these circumstances*" (Pereira and Saptawijaya, 2015: 198). Practically, part of the problem when it comes to circumscribing issue affecting machine learning and other '*elegant algorithms'* concerning ethics and justice is that thy predominantly use "*utilitarianism and deontological ethics* [for] *providing a framework to encode moral rules, typically in favour of deontological ethics, with or without referring to specific moral rules*" (2015: 199).

Considering this in more detail, deontological ethics holds that "[e]thics [is] *based on the notion of a duty, or what is right, or rights, as opposed to ethical systems based on the idea of achieving some good state of affairs* [i.e. utilitarian ethics] *or the qualities of character necessary to live well"*, i.e. virtue ethics (Blackburn, 1996: 100). In other words "*an action is considered morally good because of some characteristic of the action itself, not because the product of the action is good*", subsequently "*some acts are morally obligatory regardless of their consequences for human welfare*" (Available online: https://www.britannica.com). Notably there is then a strong link with ethics of rights, specifically negatives rights, but not proportionate to positive rights.

The problem with this being that although deontological ethics may be a good tool for "*putting the right rules and norms in place*," and to assist in "*formulating or drafting the right rules and regulations to determine right or wrong*", ethics is not as impersonal as rule based ethics would suggest (Fouché, 2006: 26).Specifically, ethics is more than deciding the rules and regulations of right and wrong, it is "*about people who act*", and therefore inherently a messy, practice as we will see in the section on talking about the problem of multiculturalism and AI ethics policies (2006: 26). Additionally, though deontological ethics it is good for delineating aspects relating to not interfering with the rights of individuals, it lacks clarity when it comes to assigning responsibility once we start talking about positive rights, in which case we usually have to refer to ethics of distributive, punitive and retributive justice to clarify or remedy.

Utilitarian ethics is defined as "[a] *general term for any view that holds that an action should be evaluated on the basis of the benefits and costs they will impose*" (Velasquez, 2006: 61). More importantly, to this article, it is "*any theory that advocates* [the] *selection of that action or policy that maximizes benefits (or minimizes cost)*" (2006: 61).

In this context the issue with *utilitarian ethics* is its inability to "*make* [any] *sense of the range of thoughts"* such as moral agency and integrity *"properly engenders*" (Blackburn, 1996: 195). Utilitarianism's basic flaw is that it has problems of measurment. Basically, some things are immeasurable, such as all foreseeable possible future consequences of AI actions, or the emotional aspects of understanding personal responsibility.

In institutions such as UNESCO's reliance on predominantly demonological and utilitarian ethics, one could argue that their machine ethics as is, is positivistic. This means they adhere to the philosophy that "*the highest or only form of knowledge is the description of sensory phenomena* […]*, so called because it* [confines] *it self to what is positively given, avoiding speculation*" (Blackburn, 1996: 294). Again, the issue relates to those parts of ethics which are not as easily measured or purely

translatable in sensory terms, such as moral indignation or '*common sense reasoning*'.

Neither of these are sufficiently adequate to when it comes to the real-world complexities of ethical dilemmas as it assumes learning from clean sets of principles that could be revised and relearned, whereas ethics is constitutionally the very process of detangling various principles, norms, value, morals, laws etc. to come to an ethical decision or new principle which may or may not be perfect, but is at least sufficiently defendable in terms of *tort* law. Conversely, it is then ethics which usually precedes the creation of laws and how we understand the fair and adequate attribution of responsibility and blame. Furthermore, because of the changing nature of the context within which principles, morals norms etc. are themselves develop and therefore fluid over time, both ethics and laws must be open to being wrong and to change, not merely recalculate.

Also, with regards to the pace of technological development versus the formation of laws, one could make the analogy that the haste for developing AI ethics, laws and policies is akin to the "*proliferation of nuclear weapons in the 1960's"* (Fung and Etienne, 2022. Available online: https://link.springer.com). However, this has meant a profusion of documents on "*AI ethical standards, as much as 84 identified by Jobin et al. [14] and 160 in Algorithm Watch's AI Ethics guidelines Global Inventory [1]*" (2022. Available online: https://link.springer.com). It is therefore unsurprising that terminology will get confusing, to which adding the layer of ethical abstraction to concepts such as AI and AI moral agency obfuscates to potential for ethicists and other stakeholders to function as they should.

Fung and Etienne are also quick to point out that even though it can be noted that among those 84 "*8 key themes across 36 of the most influential*" documents could be found, making it look as if there is a converges towards a normative approach, principally in applying Kantian deonological ethics, they are in fact "*not universal in practice*" (2022. Available online: https://link.springer.com). They point out that AI being a global phenomenon "*ethical pluralism is more about differences in which relevant questions to ask rather than different answers to a common question*" are sought, because even though people from different cultures may "*agree on a set of common principles, it does not necessarily mean that they share the same understanding of these concepts and what they entail*" (Available online: https://link.springer.com).

This is illustrated for example when comparing Chinese and EU policies involving AI and ethics, which may look the same, but "*Chinese principles emphasize the promotion of good practices* [whilst], *the EU focuses on the prevention of evil consequences*" (Available online: https://link.springer.com). The real world implications of this in practice is that "[t]*he former draws a direction for the development of AI, so that it contributes to the improvement of society* [versus] [t]*he latter* [which] *sets limitations to its uses, so that it does not happen at the expense of certain categories of people"*(Available online: https://link.springer.com). This is then merely one example of how the application of a universal deontological approach to AI and ethics fails.

This leads us neatly into the next problem with this a *priory attribution of AI* in ethics, which can be explored within the context of the multicultural dimension that a global phenomenon such as AI is. Said differently "*it is erroneous to believe that a similarity in concepts necessarily translates into a similarity*

in ethics", given that *"the same words may have different meanings from* [one] *country to another*" (Fung and Etienne, 2022. Available online: https://link.springer.com). This stands in stark contradiction to UNESCO's policy statement quoted previously of having *a systematic normative reflection* approach*, based on a holistic, comprehensive, multicultural and evolving framework,* imply coming to some universal set of *interdependent values, principles and actions.*As has been mentioned regarding the deontological approach, this also ignores the human elements of ethics as something which is not merely a conceptual set of rules, but rather also normatively the way in which people act, and in the case of Fung and Etienne, understanding people's behavior from within their own cultural philosophical perspective.

Expanding our linguistic analyses, the problem of using the improper lexicon when we talk about machine learning, robotics, AI etc. is that it obfuscates matters of real concern, notably with reference to *tort* law. Above and beyond the mere the contradictory anthropocentric approach set out in many policies wanting a *humanitarian AI*, in cases where there is confusion around topics with great socioeconomic, environmental, financial, labour etc. consequences, this uncertainty becomes a battle ground within the political arena.

This we can already see in in EU's attempts and confusion around merely defining AI (Bertuzzi, 2022. Available online: https://www.euractiv.com). In the absence of the proper use of AI and insisting on implementing the a *priory use value* of AI, which haphazardly concatenates what should be separate field of investigation, those who have to make laws and policies do so not only in haste, but inadequately. Subsequently you find that either "*the most controversial topics*" such as "*the definition of artificial intelligence (AI) itself*" gets "*pushed further down the line*", or some approximation as we have mentioned like *'functional AI'* gets used, in either case avoiding the issue (2022. Available online: https://www.euractiv.com).

Given the very serious nature of the topic is it then very disconcerting that, because of insisting on using the terms AI inappropriately, we can have a situation where for example: "*liberal and conservative MEPs are proposing a general lowering of the fines*" whilst the "*centre-left Benifei* [are] *pushing for an overall increase of the sanctions and for removing size and market share consideration from the criteria*" when it comes to the *tort* law aspect of fines and penalties (2022. Available online: https://www.euractiv.com).

In the worst-case scenarios what is happening in the legal and political vacuum created by the insistence of improper use of lexicon surrounding the topic of AI is that important institutions such as the UN or EU cannot come to a clear concise agreement on life-or-death topics such as the use of "*autonomous killer robots*" (Dawes, 2022. Available online: https://robohub.org). These are weapons on which countries such as the United Sates alone has invested roughly $18 billion, and that "*can operate independently, selecting and attacking targets without a human weighing in on those decisions*", relying heavily on *black box systems* (2022. Available online: https://robohub.org). Just how dangerous this is and the extent that it poses a threat to human rights is a topic best explicated by Dawes themself, but needless to say, much of what has been mentioned as critique could have been prevented or resolved with the accurate use and identification of the language involved.

To remedy many of the foregone criticism, it is suggested we return to 'simpler' more accurate lexicon as described in the work of Pana. This would mean that instead of employing '*functional AI'* or other versions thereof for the sake of easy of conversation or to try and circumvent some of the more technical complexities of AI and ethics, we revert to three main fields of ethics identified by Pana. These are:

1. **Ethics of Computing:** This is "*not a professional ethics, but one destined for computer and net workers with diverse professions, who process and transmit information*" (Pana, 2006: 255-256). This would include looking at "*software property protection, ensuring user identity and intimacy and* […] *the sharing and preservation of a netiquette*" (2006: 255-256)
2. **Computational Ethics:** This delineates the use of "*computers in the ethical field of philosophy for theoretical and practical moral problem solving*" (2006: 255-256). This will include areas of interpretability and explainable such as "*computer-based teaching and learning means and methods* [that] *are adopted and developed" (2*006: 255-256). It can also include "[a]*ccredited moral theories* [being] *studied by computational methods*" which could help ethicists look at the "*foundation-of-decision analysis in difficult moral problems*", where (2006: 255-256).
3. **Machine Ethics:** This involves the computer itself, including how "*the intelligent machine induces changes in the world, like humans do*" (2006: 255-256). Just as "*human activities have moral significance*", we have seen from the various AI approaches given that a *"machine with similar possibilities needs moral functions*" (2006: 255-256). An outcome of this would be "*an Artificial Ethics which will constitute a part of Artificial Philosophy*", which will prerequisite "*a strong philosophical (ontological, axiological, pragmatic and ethical) foundation"* (2006: 255-256).

Another definition for machine ethics we may use is researched focused on "*trying to find appropriate answers for the problem scope*: "*the consequences of the behaviours, which are shown by machines to humans and other machines*" (Anderson; Anderson and Armen, 2004 qtd in Kose; Cankaya and Yigit. 2018: 72). This researched is still based on, but should not be limited to the deontological "*idea of defining ethical rules to prevent* [humans] *from any possible dangerous – harmful results (especially for the humankind) caused by intelligent systems*" (Kose et al, 2018: 72)

Pana at the time also suggest **Global Information Ethics** which would take a broader perspective which is described as an *"ensemble of the above-mentioned* […] *domains of ethics*, […] *a super-structured new level of ethics, as a result of their synthesis*" (2006: 255-256). However, as we have discussed, the problem with a lexis further complicated by cultural plurality, probably means that even though there may be cross boarder objects of debate which lie beyond geographical, social, political and cultural differences, till such time that issues of *understandability* and *explainability* compared with the language implemented is resolved first, trying to approach these issues would potentially only make matters exponentially more complicated (as we have already mentioned within the arena of politics), leading to greater conflict.

Closing this section in keeping with our critical self-reflective perspective, looking at an ideas current and potential future plausibility within our socio ethical economical context, we can say that historically and contemporary AI as

propounded by different schools of thought (key to whom is its ability for *express agency* with the ability to apply *common sense reasoning*), does not exist.

Likewise, ethical policies that speak to this seem to have serious problems of explanation, generally stemming from applying *a priory positive identity* to their conception of AI, whilst at the same time coming into contradiction with AI as *express agent*. The lexicon used is therefore insufficient as it concatenates elements of machine learning and *elegant algorithm* modeling which should, for both *tort* legal and ethical consideration be separate, uniquely since consensus in matters such as the use of *black box systems* employed in tools such as killer drones, is not forthcoming. This may then require or could be remedied in part by reverting to 'older', more precise terminology, breaking the field of study up(at least for now), into its more manageable constituent parts.

Section three: The potential emergence of AI as moral agent:

Carry our methodological approach throughout, we need to look to the future potential of AI from within the historical delineation it as an *express agent*. This means looking at what *express agency* may require and its impact on the field of AI ethics. From the outset it should be stated that, unlike the *anthropocentric* view that some institutions may hold, the author is unconvinced we can fully do away with the possibility of AI, nor AI consciousness, or as *express agent*. What is more, it is possible to reach this conclusion without reverting to sentience.

What will be explored in this article, based on new research in neurobiology- and -science, especially the uncomfortable fact that much of our brains do work like programs (markedly corresponding to free will[18], and as this does not contradict classes of exiting as is the case of sentience), we can say that: Though AI may still be very far off and still a concept in development, the likelihood of the latter exists.

Additionally (from a philosophical nouminal existential perspective), findings about how some molds behave (solving labyrinths better than human engineers[19]), or chimps who do better in some forms of memory recall than their human counterparts, questions remain as to:

> At what level do we speak of being in the presence of consciousness and learned behaviors?

Likewise, in terms of *computational* – and -*machine ethics*, work being done on how machines learn may help with the diagnosis's illnesses such as Autism[20], as well as the possibility of learning more about consciousness itself from something like AI. Principally: Considering the historical fact that the nature of consciousness is a philosophical question as old as philosophy itself (from Hermeneutics, the ontological proof of god, to the delineation between phenomenology and existentialism and beyond), these question must surely take on a different dimension when the physical time limitation of mortality is no longer a consideration, as would be the case for an AI. From merely this point alone already it is safe to say that philosophically,

[18] As discussed later in this section.

[19] See: "*How Brainless Slime Molds Redefine Intelligence"* (Jabr. 2012. Available online: https://www.scientificamerican.com), as well as: "*Chimps outperform humans at memory task"* (Hooper. 2007. Available online: *https://www.newscientist.com*).

[20] See: New AI-Driven Algorithm Can Detect Autism in Brain "Fingerprints" (Hadhazy, 2022. Available online: https://hai.stanford.edu/news/

when talking about AI consciousness it would probably be a different kind of consciousness.

There is however a more profound reason for this relating back to contemporary AI's inability for *common sense reasoning* or *moral creativity* which necessitates its development towards being able to hold *superposition*:

With the application of something akin to *Qbits* and the computation power that lies behind it, if AI consciousness did emerge having Quantum *superposition* (which would lend it the ability for ethical and moral reasoning), such a consciousness would be on a different plane s to ours[21]. Said differently "[m]*achine ethics can/will be of the highest quality because it will be derived from the sciences, modelled by techniques and accomplished by technologies*" which will exceed biological limitations in most respects (Pana, 2006: 254). This is already transpiring to some degree in the field of "s*uperintelligence*" research which is "*based on the idea of an intelligence type making intelligent machines to "surpass human brain in general intelligence"* (Bostrom, 2014 qtd in in Kose; Cankaya and Yigit, 2018:73)

Another way in which we could understand AI becoming conscious or at least an expression thereof, is through the idea of memes as living concepts. The argument follows that: Memes are like cognitive viruses which 'live' vicariously through humans as they spread from one person to another (Blackmore, 2013. Available online: https://www.philosophytalk.org). From this perspective it could be contended that the concept of AI has spread through our consciousness, infecting our minds, driving our need for it to exist, but in fact like a virus it can only straddle the edges of being alive conceptually and not physically. Conversely, it could be stated that the very existence of the concept AI is an expression of consciousness itself trying to understand the universe through break the physical limitations of human existence by using the latter as a host to find solutions to the problems of mortality. However this takes us into the fields of metaphysics which, as was stated from the outset we will steer clear of for the sake of avoiding class confusion as much as possible.

There is another alternative which stays within the epistemic world in the studies of George Lakoff who looked at the physical effects of metaphors on our cognitive processes. Using the observations of Jerome Fieldman on "*mirror neurons*", which is the fact that the same nodes of neurons fire when we say the word 'grasp' than when we see someone grasping something, they developed the idea of "*simulation semantics*" (Lakoff, 2009: 3. Available online: http://www.neurohumanitiestudies.eu).

They argue that "*if you cannot imagine someone picking up a glass, you can't understand the meaning of "Someone picked up a glass*", subsequently it follows that "*for meanings of physical concepts, meaning is*

[21] Though we may be some way off from the emergence of an AI due to the problems mentioned, at least in the field of quantum computing big strides have been made with the first quantum circuit being built this year "*in which each atom has multiple quantum states"*, i.e. superposition (Nelson, 2022. Available online: https://www.sciencealert.com). To illustrate how much more powerful quantum computing is, if for example one wanted "*to create a simulation of the penicillin molecule with 41 atoms, a classical computer would need* [ten to the power of eighty six] *transistors, which is "more transistors than there are atoms in the observable universe*", whilst a quantum computer would *"only require a processor with 286 qubits*" (2022. Available online: https://www.sciencealert.com).

mental simulation" (2009: 3). If then AI as an *elegant algorithm* is a metaphorical expression of our own consciousness, it would make sense for us to want to will its physical existence bearing in mind "*all mental simulation is embodied,* [as] *it uses the same neural substrate used for action, perception, emotion, etc"* (2009:3).

Furthermore, as we have seen intrinsic to the very definition of AI is the prerequisite for *explicit agency* separate from human interference. We could say that AI lends its physical consciousness' by the mere notions of our imagining its existence, though the problems of this have already been discussed previously on the subject of the denoting responsibility and so, though we could argue that consciousness' could arise from us, the very idea of AI stemming from our own hands comes into internal conflict with AI's *explicit agency*. However, within the confines of a *ghost in the machine*-like scenario (or a more evolutionary perspective), the case could be made for the physical manifestation of our metaphorical understanding of AI that would have its genesis from such a wide source of human cognition so as to make any one particular person or party responsible impossible and negligible. This would also akin to our own cognitive development which requires both vicarious and enactive learning as to make the differentiation between what was inherent or not so complex that to do so would again boarder on the metaphysical, i.e. falling back on the notion of having or not having a soul.

Contemporary neuroscience adds to the discussions through its study on the concept of free will. This is because it looks at:

> Are we merely like robot's reactionary machines to external stimuli, or if there is more to our thinking process?

In philosophical terms it relates back to the problem of mind body duality (which we will briefly look at later)

The question of free will in neuroscience begins in the 1960's with the study done by "*Kornhuber and Deecke*" which resulted in what they termed the "*Bereitschaftspotential",* where it seemed that the brain readied itself before making a decision (Gholipour, 2019. Available online: https://www.theatlantic.com). This was followed up by the work of Liebet which focused on "*the allegedly unconscious intentions taking place in decisions regarded as free and voluntary*" (Lavazza, 2016. Available online: https://www.frontiersin.org). In this study it was found that "*the Bereitschaftspotential started to rise about 500 milliseconds before* […] *participants performed an action",* and only reported *"their decision to take that action* […] *about 150 milliseconds beforehand*" (Gholipour, 2019. Available online: https://www.theatlantic.com). Suggesting that "[t]*he brain* […] *'decides' to initiate the act" before a person is* […] *aware"* of making the decision (2019. Available online: https://www.theatlantic.com). If this is the case then what we call consciousness, which is ostensibly a creation of the mind, is simply an accumulation of mechanical responses and the difference between us a machine thinking would be very little.

However, since then research conducted 2010 by Aaron Schurger found that this was an over simplistic view. They found that what had actually been measured was the background ebb and flow of our brain's natural neural activity (2019. Available online: https://www.theatlantic.com). As is,

the contemporary scientific problem around free will can be summarised as follows:

> "*the internally generated brain activity has to do both with the stochastic noise and with the history of the subject's choices. On the one hand, the stochastic noise comes both from the configuration that the brain has on average as a result of evolution (adaptive significance) and from individual development, resulting from random processes and environmental influences. On the other hand, the history of the choices is derived from the same process (in part stochastic* [and deterministic]*) that […] have just described*" (Lavazza, 2016. Available online: https://www.frontiersin.org).

In other words, free will is at a paradoxical impasse similar to that of the double hermeneutic problem whereby: The mere act of observation of behaviour changes behaviour and so one has a problem in measurement. Similarly, one could draw analogies to explicit agency in AI where it could be said that the current view holds that: The very moment something like AI acts in a way that we could consider it to have *explicit agency*, it seizes to have that agency because such behaviour would have its genesis in human agency. A catch twenty-two situation where in fact the only way forward for an AI to have *explicit agency* would be to prove unequivocal *superiorintelligence*.

What is more "[m]*odern neuroscience tells us that we are completely unaware of most brain activity,* [and] *that unconscious processing influences behaviour"* (Costandi, 2022: Available online: https://bigthink.com). This follows studies on the unconscious mind which track similar lines of reasoning than what was mentioned in reference to neural nodes, *mirror neurons* and *simulation semantics* which state that mental simulations are *embodied.* In recent studies it has been shown that "*unconsciously processed visual information is distributed to a wider network of brain regions involved in higher-order cognitive tasks*", subsequently it was found that "*unconscious processing contained meaningful information about the images, which became accessible to higher-level stages of processing*" (2022. Available online: https://bigthink.com). The implications being that there is a greater overlap between conscious and unconscious mind than previously thought which requires a revision of the "*Global workspace theory of consciousness*" that holds that of "*information are processed in their relevant local domains, and only enter conscious awareness if they are first received, and then shared by, the central hub*" (2022. Available online: https://bigthink.com)

Comparing what neurobiology has told us, it seems that when it comes to AI and the possibility of emergent consciousness the fact is also that "[t]*here is real causation going on between various units of brain activity precisely mirroring patterns of causation between the neurons*", making the aforementioned emergence theoretically plausable (Nath and Sahu, 2020: 105). Furthermore, in terms of the physical technology there is also the developments from the university of Princeton who are building an "*artificial brain*" (Available online: https://www.linkedin.com). All this simply means that though we may dispense easily with AI and sentience, it not so simple when we seriously deliberate it within the context of consciousness.

This is a far less controversial statement to make considering statements already made vis-à-vis the historical philosophical nature of consciousness. As early as Socrates the problem of mind body duality has persisted with some speculating that the latter is like the ether which we tap into, and other as seeing

consciousness as an emergent property stemming from our participation and reasoning about the world around us. One of the most extreme of these views is called "*Epiphenomenalism*" which states *"that our conscious minds serve no role in affecting the physical world property*" (Thomson, 2020. Available online: https://bigthink.com). It is argued that "*our thoughts are a causally irrelevant byproduct of physical processes that are occurring inside of our brains*", as such our mind from which our consciousness stems is a mechanistic reaction to the world (2020. Online). Furthermore, though the question of AI consciousness is often given as a key concern when talking about its ability to hold *explicit agency*, there are some that we have already mentioned in section one who would argue that phenomenologically, based on recent neurobiological finding, its significance is questionable (Behdadi and Munthe. 2020: 197).

What we can gain from this is that we find ourselves on shaky foundations on the very same limitations we set for an AI to have *explicit agency* ourselves. If our own free will which determines matter of mitigating circumstances and punitive justice in *tort law* and the consciousness on which we build intent to act can be brought into question, who are we to hold another to such prerequisites? We may argue it is fair because the stakes are so high, but then we also may need to reconsider the consequences of the alternative, which we will in what is to follow. In any event, the outcomes to these issues of consciousness and free will concerning ourselves and AI have two potential outcomes: On the one hand one has the acceptance of moral agency (and in this case along with it AI), with the possibility of ethical behaviour, and on the other the type of moral nihilism best described by *Epiphenomenalism.* Accordingly, if we accept the latter there would be very little need to concern oneself with ethics and deliberating the effects of one's actions on others.

None the less, what the accumulate effect of the aforementioned sections seek to elucidate was that though the criteria for AI to hold *explicit moral agency* may not require consciousness (at least not as we know or can even define for ourselves), and that the bar for it to be recognised as such practicing a kind of *superpositioning* and reflexive thinking we would expect from humans, requiring as mentioned that the bar to be set incredibly high, it's very existence is still an future epistemological plausibility. This is all we needed to establish to fulfil the methodological prerequisites of Adorno's negative dialectics. The question which then arises is: What is the implications of this plausibility with regards to the practice of AI ethics?

To answers this, if we look to the different perspectives of AI covered earlier, though consciousness would merely be a prerequisite both views hold that AI should have some form of social community and 'belief' or intentionality within which it would function. To this end we can already see the beginnings of such a social structure as part of what is called Distributed Artificial Intelligence (DAI).

DAI is a precursor to Multi-Agent systems and uses distributive approach to solving problems that are especially complex. More specifically DAI is *"a class of technologies and methods that span from swarm intelligence to multi-agent technologies and that basically concerns the development of distributed solutions for a specific problem*" (Corea, 2009. Available online: https://francesco-ai.medium.com)

In accordance to Pana and others AI prerequisites, it can already be said that the "*collective realm, norms and moral emergence has been studied computationally, using the techniques of Evolutionary Game Theory*" (Pereira and Saptawijaya, 2015: 198). In fact, if we take seriously that an AI prerequisite *superposition* would probably function on a higher cognitive ability than most humans and contextualize within the realm of *DAI* research, has already found that with "*the introduction of cognitive capabilities, such as intention recognition, commitment, and apology, separately and jointly* [will] *reinforce the emergence of cooperation in the population* (2015: 198).

What is more, based on logic programming which already exists it has been found that "*modeling moral cognition in individuals* […] *within a networked population shall allow them to fine tune game strategies, and in turn may lead to the evolution of high levels of cooperation*" which could then lead to understanding "*the emergent behavior of ethical agents in groups* […] *and their swarms*" (2015: 198). A concern here being that if AI required *Qbits* to perform functions of *reflexivity* to achieve *explicit agency*, such agent(s) may 'speak' a language only '*the swarm*' understand. In effect creating a self-enclosed community that has the ability to communicate with us, but which we may have no recourse to communicate with in their own lexicon.

As for *intentionality* or '*faith*' it is not completely inconceivable that *black box* programming can be seen as analogous to our unconscious mind where in "[c]*onsciousness is but the tip of an iceberg poking up from the waters of our brain, with the vast bulk of its capacity – the subconscious mind – hidden underneath the surface*" (Adams, 2010. Available online: https://www.stateofdigital.com). Just as with human then such an AI with *Qbit* computing that can exist in a community of *DAI* could potentially have unique elements that are hidden from others and as with humans, acting as motivations or intent for their behaviours and choices.

Having said this, with regards to the topic of AI and ethics itself: If we take all of this into consideration what it would actually take for and AI to practice '*reasoning*' and emerge into fruition, the sheer computing power it would require to resolve the issues related to *specific* and *universal* sets of knowledge will, as has been argued [previously probably mean having to implement quantum computing[22]. The subsequent actual issues that would present ethics, ifread against the societal *DAI* and *black box unconsciousness intentionality* set here, would bring into question whether of not our anthropocentric ethical conceptions will even apply to AI? The reasons being:

- Firstly, why should a being that would work on a higher plane of cognitive processing than any human want to adhere to human standards of ethics?
- Secondly, if we are talking about *DAI*, they would be able to create their own society with their own mores, morals norms and values, circumventing the need to adhere to human conceptions of the latter (especially if we add point one)

[22] See footnote 14 about Qbits superposition. The reason why quantum computing will probably be required is because quantum bits are more stochastic, in that they can hold more than one position at the same time to cope with variability, not merely deterministicly holding either only (1) or a (0), and therefore possesses the possibility to simultaneously hold specific as well as universal bits of data.

- Thirdly, if we are talking about Ambient AI[23] relative to point one and that an AI could in fact be in a disembodied form, it is quite possible such a being would already know the greatest threat to its own existence are humans and may never decide to reveal itself and only work behind the scenes to ensure its own existence, in which case we would simply become its senses, just like our eyes, ears, nose, tongue and skin merely do our bidding.
- Lastly, and more worryingly AI could potentially function on such a level beyond our own human limitation, such as limitations of human language and the difficulties of defining (has already been illustrated and talked about by hermeneutic philosopher, Adorno, Wittgenstein and many other), that it may consider itself beyond ethics and unable to make unethical choices. Said differently, it could possible view its 'mistakes' as mere miss calculations which in and of itself is something very different from taking responsibility for ethical conduct.

In conclusion:

The matter of ethics and AI as it is currently being approached is being hampered by its ideological and psychological conflicts between wanting to *'will it into existence'*, versus not believing it will ever emerge. This not only leads to bending the historical accounts of AI to fit their own sociopolitical aims (problematising the conception of a singular definition of AI that could speak to the potential legal, political, economic and environmental threats it could possibly pose), but also utterly leaving out 'true AI' as in keeping with its ontological conception. As such we can surmise that Ethics about AI as it being approached now is an oxymoron, i.e. a concept that is made up of contradictory or incongruous elements.

This in turn poses a threat in and of its own as it blinds us to the potential that: If such a thing did come to fruition it may function in ways which would make our own anthropocentric approach to AI and ethics mooted. Such a conceptual blind spot may mean missing a potentially more dangerous disposition around discussing AI and ethics. Contemporarily this may mean having to revert to a 'simpler' lexicon which talks about learning algorithms instead of AI that could potentially resolve problems of definitions which have legal ramifications involving accountability, hampering current Ai ethics policy and legal deliberations.

Conversely, to not to do so would be unethical given that , though some may say that certain algorithms have a level of autonomy where they are making 'decision' and 'taking actions' which their programmers could not foresee (as for example in the case of "*black box*" type programming), to then go an try an mitigate the responsibility of those companies and programmers for the consequences taken by these 'elegant arithmetic objects' would be to ignore the ethical principle of *special offices*.

Consequently, above and beyond the fact that the continued use of AI is an attempt to *a priory* 'positive identity' (being subverted by anthropomorphic of AI, relegating it to simple Robotics and machine learning), insisting on using the idea of AI and ethics would be to

[23] Ambient AI or AmI is "[t]he *ability of technology to take decisions and act on our behalf taking into consideration our preferences based on the data available to it from all the connected sensors and systems surrounding the user*" (Doddavula, n.d. Available online: https://www.infosys.com).

obfuscate the key principle of personal responsibility central to ethics, by trying to placate or ignore the fact that regardless of the types of consequences the holder of a *special office* cannot rely on either mitigating arguments of ignorance or inability to act. In truth not even considering the latter, the very fact that we have terms such as '*black box*' and 'autonomous drones' means the variable of unpredictability has been recognized. What has not happened in concert is that the full potential for such variables implications has not been taken into consideration, as technological expansion has happened faster than policy and legal frameworks could. Underscoring yet again, that adding a layer of pseudo complexity by applying a term which is unfit for purpose does not serve ethical or legal purposes and hampers proper risk management in the future.

Bibliography:

1. Adorno, T. W. (1973). *Negative Dialectics (1966). Translated from the German by Ashton, E. B. London*: Routledge & Paul Ltd

2. Anon. (2022). *A European approach to artificial intelligence*. Available online from: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (Accessed 21 August 2022)

3. Anon. (2021). *Daft text of the recommendation on the ethics of Artificial Intelligence (PARA 1-25; Text considered in the 1st session)*. Available online from: https://unesdoc.unesco.org/ark:/48223/pf0000377881 (Accessed 9 July 2022)

4. Adams, B. (2010). *Google, the internet and artificial intelligence*. Available online from: https://www.stateofdigital.com/google-the-internet-and-artificial-intelligence/ (Accessed 8 June 2022)

5. Anderson, M; Anderson, S. L and Armen, C. (2004). Towards machine ethics*. In Proceedings of the AOTP'04-The AAAI-04 Workshop on Agent Organizations: Theory and Practic*e

6. Anderson, M and Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI magazine*. 28(4): 15-27

7. Anderson, M and Anderson, S. L. (2011). *Machine Ethics*. London. Cambridge University Press

8. Badoius, A. (2006). *Briefings on existence: A short treaties on transitory ontology. Translated from the French by Madarasz, N*. New York. SUNY Press

9. Behdadi, D and Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*. 30 : 195-218

10. Bertuzzi, L. (2022). *AI regulation filled with thousands of amendments in the European Parliament*. Available online from: https://www.euractiv.com/section/digital/news/ai-regulation-filled-with-thousands-of-amendments-in-the-european-parliament/ (Accesses 17 June 2022)

11. Blackburn. S. (1996). *Oxford Dictionary of Philosophy*. Oxford: Oxford University Press.

12. Blackmore, S. (2013). *Memes: Viruses of the mind*? Available online from: https://www.philosophytalk.org/shows/memes-viruses-mind (Accessed 17 August 2022)

13. Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. London. Oxford University Press

14. *Building an artificial brain*. Available online from:

https://www.linkedin.com/feed/update/urn:li:activity:6948264380752896001/?utm_source=linkedin_share&utm_medium=ios_app (Accessed 4 July 2022)

15. Corea, F. (2009). *Distributed Artificial Intelligence. A primer on Multi-Agent Systems, Agent-Based Modeling, and Swarm Intelligence.* Available online from: https://francesco-ai.medium.com/distributed-artificial-intelligence-3e3491e0771c (Accessed 18 August 2022)

16. Costandi, M. (2022). *Neuroscience research triggers revision of a leading theory of consciousness*. Available online from: https://bigthink.com/neuropsych/revision-leading-theory-consciousness/ (Accessed 20 July 2022)

17. Dawes, J. (2022). *UN fails to agree on 'killer robot' ban as nations pour billions into autonomous weapons research*. Available online from: https://robohub.org/un-fails-to-agree-on-killer-robot-ban-as-nations-pour-billions-into-autonomous-weapons-research/ (Accessed 15 July 2022)

18. *Deontological ethics*. (2020). Available online from: https://www.britannica.com/topic/deontological-ethics (Accessed 7 July 2022)

19. Dickson, B. (2022). *Large language models have a reasoning problem*. Available online from: https://bdtechtalks.com/2022/06/27/large-language-models-logical-reasoning/amp/ (Accessed 30 June 2022)

20. Doddavula, S. K.(n.d.). *Living with Ambient Intelligence: So at home with technology.* Available online from: https://www.infosys.com/insights/ai-automation/ambient-intelligence.html (Accessed 19 July 2023)

21. Floridi, L; Cowls, J; Beltrametti, M; Chatila, R; Chazerand, P; Dignum, V; Luetge, C; Madelin, R; Pagallo, U; Rossi, F; Schafer, B; Valcke, P and Vayena, E. (2018). *AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations*. Atomium: European Institute for Science, Media and Democracy

22. Fouché, J. B. (2006). *Trust and business. An inquiry into the functioning of trust in business. Masters of Business Administration*. Cape Town: Graduate School of Business of the Stellenbosch University

23. Fung, P and Etienne, H. (2011). Confucius, cyberpunk and Mr. Science: comparing AI ethics principles between China and the EU. *AI and Ethics*. Available online from: https://link.springer.com/article/10.1007/s43681-022-00180-6 (Accessed 12 July 2022)

24. Garcia, E.V. (2021). *UNESCO's Recommendation on the Ethics of AI: why it matters and what to expect from it*. Available Online from: https://thegoodai.co/2021/11/24/unescos-recommendation-on-the-ethics-of-ai-why-it-matters-and-what-to-expect-from-it/ (Accessed 1 July 2022)

25. Gholipour, G. 2019. *A famous argument against free will has been debunked*. Available online from: https://www.theatlantic.com/health/archive/2019/09/free-will-bereitschaftspotential/597736/ (Accessed 17 July 2022)

26. Hadhazy, A. (2022). *New AI-Driven Algorithm Can Detect Autism in Brain*

"Fingerprints". Available online from: https://hai.stanford.edu/news/new-ai-driven-algorithm-can-detect-autism-brain-fingerprints#:~:text=and%20Cognitive%20Science-,New%20AI%2DDriven%20Algorithm%20Can%20Detect%20Autism%20in%20Brain%20%E2%80%9CFingerprints,timelier%20interventions%20and%20better%20outcomes.&text=Stanford%20researchers%20have%20developed%20an,by%20looking%20at%20brain%20scans. (Accessed 18 July 2023)

27. Hagendorff, T. (2022) *AI ethics and its pitfalls: not living up to its own standards?*. AI Ethics. Available online from: https://doi.org/10.1007/s43681-022-00173-5 (Accessed 6 June 2022)

28. Hofstadter, D. R. (1999). *Gödel, Escher, Bach: an eternal golden braid*. New York. Basic Books Inc.

29. Hooper, R. (2007). *Chimps outperform humans at memory task*. Available online from: https://www.newscientist.com/article/dn12993-chimps-outperform-humans-at-memory-task/ (Accessed 18 July 2023

30. Hughes, J. (2010). Contradictions from the enlightenment roots of transhumanism. *Journal of Medicine and Philosophy*. 35(6): 622-640

31. Jabr. F. (2012). *How Brainless Slime Molds Redefine Intelligence*. Available online from: https://www.scientificamerican.com/article/brainless-slime-molds/

32. Johnson, D. (2006). Computer systems: Moral entities but not moral agents. *Ethics and information technology*. 8(4): 195-204

33. Johnson, K. (2022). *LaMDA and the Sentient AI Trap*. Available online from: https://www-wired-com.cdn.ampproject.org/c/s/www.wired.com/story/lamda-sentient-ai-bias-google-blake-lemoine/amp (Accessed 16 June 2022)

34. Keyser, J. (2009). *A Critique of compliance. Towards implementing a critical self-reflective perspective*. Stellenbosch: Stellenbosch University Publishers

35. Kose, U; Cankaya, I. A and Yigit, T. (2018). Ethics and Safety in the Future of Artificial Intelligence: Remarkable Issues. *International Journal of Engineering Science and Application*. 2(2): 71-76

36. LaFollette, H. (2002). *Ethics in Practice: An Anthology*. Oxford. Massachusetts: Published by Blackwell Publishing Ltd

37. Lakoff, G. (2009). *The Neural Theory of Metaphor*. Available online from: http://www.neurohumanitiestudies.eu/archivio/SSRN-id1437794The_Neural_Theory_of_Metaphor.pdf (Accessed 18 August 2022)

38. Lavazza, A. 2016. *Free will and neuroscience: From explaining freedom away to new ways of operationalizing and measuring it*. Available online from: https://www.frontiersin.org/articles/10.3389/fnhum.2016.00262/full (Accessed 18 July 2022)

39. Linardatos, P; Papastefanopoulos, V and Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*. 23(18): 1-45

40. Marcus, G. (2022). Artificial general intelligence is not as imminent as you might think. Available online from:

https://www-scientificamerican-com.cdn.ampproject.org/c/s/www.scientificamerican.com/article/artificial-general-intelligence-is-not-as-imminent-as-you-might-think1/?amp=true (Accessed 8 July 2022)

41. Nath, R and Sahu, V. (2020). The problem of machine ethics in artificial intelligence. *AI and Society*. 35 (1):103-111

42. Nelson, F. (2022). *A huge step forward in Quantum Computing was just announced: The first-ever Quantum Circuit.* Available online from: https://www.sciencealert.com/a-huge-step-forward-in-quantum-computing-was-just-announced-the-first-ever-quantum-circuit/amp (Accessed 8 July 2022)

43. Pana, L. (2006). Artificial Intelligence and Moral intelligence. *triple.C*. 4(2): 254-264

44. Pereira and Saptawijaya. (2015). 'Bridging Two Realms of Machine Ethics' in White, J and Searle, R. *Rethinking Machine Ethics in the Age of Ubiquitous Technology*. Pennsylvania. IGI Global Publishers: 197 -224

45. Pretz, K. (2021). *Michael I. Jordan explains why today's artificial-intelligence systems aren't actually intelligent*. Available online from: https://spectrum-ieee-org.cdn.ampproject.org/c/s/spectrum.ieee.org/amp/stop-calling-everything-ai-machinelearning-pioneer-says-2652904044 (Accessed 31 July 2021)

46. Rebber, A. S. (1985). *Dictionary of Psychology*. London: Penguin Books Ltd.

47. *Reification.* (n.d.). Available online from: https://dictionary.cambridge.org/dictionary/english/reification (Accessed 3 July 2023)

48. Singer, P. (1975). *"All Animals are Equal," in Animal Liberation: A New Ethics for our Treatment of Animals*. New York. Random House: pp. 1–22.

49. Thomas, E. L. (2004). *Emmanuel Levinas: Ethics, Justice, and the Human Beyond Being*. London. Routledge.

50. Thomson, J. (2020). *Epiphenomenalism: one of the most disturbing ideas in philosophy*. Available online from: https://bigthink.com/thinking/epiphenomenalism-mind-body-problem-dualism/ (Accessed 14 July 2020)

51. Velasquez, M. G. (2006). *Business Ethics: Concepts and Cases. Sixth Edition*. New Jersey: Prentice-Hall, Inc. (now known as Pearson Education, Inc.).

52. Wagoner, R. (2004). *Artificial Intelligence*. Available online from: http://people.moreheadstate.edu/fs/r.wagoner/IET600/projects/project5.pdf (Accessed 4 June 2013)

53. *What is tort law?* (n.d.) Available online from: https://www.thelawyerportal.com/careers/areas-of-law/areas-legal-practice/tort-law-guide/#:~:text=What%20Is%20Tort%20Law%3F,different%20types%20of%20legal%20issues. (Accessed 14 July 2022)

54. Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus.* New York. Harcourt, Brace and Company, Inc.

55. Zuidervaart, L. (2015). *Theodor W. Adorno*. Available online from:

http://plato.stanford.edu/entries/adorno/ (Accessed 28 June 2022)