Ethics in Artificial Intelligence

A Multidisciplinary Approach

Edited by

Steven S. Gouveia

Ethics in Artificial Intelligence: A Multidisciplinary Approach

Edited by Steven S. Gouveia

2025

Ethics International Press, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (Hardback): 978-1-83711-091-9
ISBN (Ebook): 978-1-83711-092-6

Table of Contents

Introduction

The world is rapidly changing entering, more and more, to the Age of Artificial Intelligence (Gouveia, 2020): we are living a technological *renaissance* that is redesigning not only how we live and interact, but also the very core of how we define essential societal concepts such as knowledge, morality, and even humanity itself. Far away from the old times when Artificial Intelligence (AI) was a matter of ideal speculations, more and more we are, as humans, embedded in technological systems that are present in legal, medical, military, cultural or educational practices, as you will see in this inter-disciplinary book.

These transformations have, necessarily, deep philosophical and ethical consequences: how should we interact with systems that are more and more autonomous in their acting? Is it possible to align human values in these technologies? How war is changing with the incorporation of AI in different military stages?

The goal of this volume titled *Ethics of Artificial Intelligence: A Multidisciplinary Perspective* is to discuss these and other relevant questions that cross several disciplines and require the expertise of many different leading areas. The main intuition behind this project is to avoid offering both utopian fears or technocratic passions and provide, instead, sustained and critical perspectives on conceptual and practical issues that follow into the AI Ethics discipline.

Also, there is true commitment to a multidimensional inquiry, where the contributors of this volume address AI in contact with other disciplinary boundaries – such as philosophy, law, cognitive science, computer science or linguistics –, to provide a more accurate image of the kind of impact AI will have in the moral terrains we will occupy in the next decades.

Therefore, you will not find in this book any kind of uniform thinking but, on the contrary, a pluralistic dialogue of different methods and disciplines that will enrich (instead of simplifying) the ethical conversation needed to provide sufficient ethical standards to deal with Artificial Intelligence in the 21st Century.

***

The volume begins with a contribution by Antonino Drago with a penetrating critique of AI's epistemic architecture. His chapter, *Artificial Intelligence as a Dramatic Challenge,* argues that AI – deep-rooted in classical logic – cannot simulate the intuitive, human kind of reasoning that supports scientific discovery. He advances a bold proposal: replace the Turing Test with a new metric grounded on the recognition of doubly negated propositions, revealing a frontier AI systems cannot cross without losing their logical foundation.

Next, Cosimo Palma in the chapter *No AGI without XI, no XI without I* takes us into the kingdom of futurology and speculative ethics by using methods like the Futures Wheel and Horizon Scanning, exploring some paradoxes and borderline scenarios that emerge as AI evolves toward general intelligence. His main claim is that ethical reasoning about AGI ought to involve existential reflection and imaginative foresight, not purely technical regulations.

Denis Coitinho follows with a chapter titled *Artificial Intelligence and Reflective Equilibrium*, where he applies the Rawlsian method of moral justification to machine ethics in general, and driving autonomous vehicles in particular. He proposes that moral consistency in AI design requires not just the traditional deontological rules or consequentialist calculus but, and more importantly, a deliberative, reflective, balance that echoes our plural moral intuitions.

Following, a chapter titled *The Ethics of Social Robotic* by Germán Massaguer Gómez investigates the emotional and relational dynamics of human-robot interaction: warning against emotional manipulation by social robots, it is argued that ethical frameworks that account for vulnerability and dependency should be considered, particularly in cases where children, the elderly, and the cognitively impaired are involved.

In Marta Mendonça's *Artificial Intelligence and Personality*, the focus shifts to metaphysical terrain by questioning several philosophical concepts – such as *intelligence*, *ethics* and *personality* – in relation to AI agents, raising questions about legal responsibility, intentionality, and the boundary between human and artificial personhood.

Parashar Das follows by adding depth to the metaphysical debate with the chapter *Agency in Artificial Intelligence Systems*, exploring whether AI systems can truly be said to possess agency, or whether they are better conceptualized as tools without moral standing. Drawing from analytic philosophy, he articulates the conditions under which agency might emerge or be attributed.

In *Ethics, Artificial Intelligence and Machine Translation in Literary Translation*, Daniela A. Rodrigues and Ana R. Pina tackle the intersection of AI and aesthetics. They examine the ethical challenges posed by AI-assisted literary translation, such as authorial intent, semantic fidelity, and quality degradation, focusing on the many ethical consequences of adopting this specific application of AI in translation.

Next, Eleni Angelou, in *Artificial Intelligence Safety as an Emerging Paradigm*, frames AI safety not as a technological add-on, but as a *Kuhnian* paradigm shift that should be widely adopted, exploring how safety concerns – both short-term and existential – are reshaping regulatory norms and forcing developers to reimagine risk, uncertainty, and control.

Turning to the practical implementation of ethical frameworks, Koray Karaca and Pablo Muñoz in *Managing the Inductive Risk of Ethical Decision-Making with Autonomous Systems*, argue that cost-sensitive optimization is at the heart of ethical risk management. Their approach draws from philosophy of science and decision theory to show how ethical considerations must be embedded in the very architecture of machine learning systems.

In *Neurotechnologies and New Processes of Subjectivation*, Luiza Cunha and Marcela Castanheira draw from Michel Foucault to examine the normativity of brain-computer interfaces, arguing that such technologies reshape subjectivity, autonomy, and embodiment, calling for a "genealogical" ethics that resists the reduction of the self to data streams.

Focusing on the same technology, Steven S. Gouveia, in *Brain-Machine Interfaces and Artificial Intelligence*, extends this ethical analysis by assessing a recent development of traditional BMIs that are AI-driven. Offering a critical reflection on how this AI-variant of brain-machine interfaces raise new challenges and opportunities to our understanding of cognition, intention, and the body, and what this means for future ethics.

Next, Yinchun Wang, in *Ethical Risks of Generative Artificial Intelligence and Its Governance*, uses China's experience with generative AI to illustrate the persistence of old ethical problems in new technological forms, providing a policy-oriented analysis that blends legal insight with cultural critique.

Finally, Roman V. Yampolskiy closes the volume with *On Monitorability of AI*, a philosophical-empirical investigation into whether AI systems can be monitored in principle, argue for the impossibility of consistently predict AI capabilities that can arise in the future, considering the consequences for AI safety research in general and propose potential strategies to overcome these limitations.

Together, these chapters compose a multi-lens examination of some of the most urgent challenges of our era. Rethinking AI Ethics requires us more than just asking what kind of machines we want to develop and use but, more importantly, what kind of humans we want to become during this transforming process. Hopefully, this book will offer neither final answers nor false comfort, but something more essential: ethical rigour, intellectual diversity, and a compass for steering the uncertain terrain forward.

Steven S. Gouveia
02 | May | 2025
Porto

Chapter 1
Artificial Intelligence as a Dramatic Challenge: What of Human Reason AI Cannot Implement

Antonino Drago[1]

The present paper characterizes within a historical and philosophical framework the abrupt challenge represented by the introduction of AI in social life is. It is then presented an alternative way of organizing a scientific theory that since centuries ago was applied by many prominent scientists, even the most revolutionary ones in the history of science. In this organization the way of arguing belongs to intuitionist logic, which in particular makes use of doubly negated propositions, *ad absurdum* arguments and the principle of sufficient reason. It is shown that AI, managed by classical logic, cannot implement any of these three features of such a theoretical organization. AI can only simulate in a parallel way a human reasoning based on them. This result implies that Turing test may be overcome. Rather a new test of intelligence is suggested; it is based on the use of intuitionist logic.

Keywords: Viewpoints on AI, mankind consciousness, alternative scientific organization, Turing test, new test of intelligence.

[1] University "Federico II" of Naples, via Benvenuti 3, Calci 56011, Italy – drago@unina.it.

1. Introduction: The Distressing Situation of Mankind Self-consciousness in the Time of AI Birth

The history teaches us that the growth of the lethal consequences of one person's act is impressive. In ancient times each time a person acted with lethal arms could do evil to few humans; the consequences of his actions were very limited in space and time. Today the launching of a nuclear bomb on a town is capable to give death to one million and more persons. However, in this case the ultimate capability and responsibility of this act pertains of a state, because the preparation of such weapon is very complex and difficult. Instead at present time Artificial intelligence (AI) allows each person to provoke through usual computer disruptive damages to mankind. Therefore, the capability of e.g. genocide has been lowered from states, which are great and rational collective social groups, to little groups and at last to a single person. This is an impressive point: we are not prepared to collectively manage so great capability of destruction as that offered by AI to even a single person.

However, the most disconcerting aspect of the present AI problem is its unforeseen occurrence. "But shouldn't someone have warned us?" is the famous question of queen of England when the subprime economic crisis occurred in the year 2008. In other words, since we are living in a modern society, where the life can rely on a lot of great advantages and privileges, which all together model a semi-god life, how is it possible that such a radical economic crisis could occur abruptly?

The same may be reiterated about the happening of the AI problem. In a parallel way to the previous queen's question one can ask: why in a so advanced stage of the development of our scientific society no one warned us that a relatively young science, computer science, could generate a so painful crisis? We are almost absolute masters of the world through science and technology. Yet, since past half a century AI became even more powerful than human intellectual faculties so that at present AI appears as uncontrollable in its capability of provoking social disruptions. In particular, we abruptly have met a dramatic question: can our most advanced instrument of power, AI, turn against us and reduce us to servitude? Can we avoid a dominion of mankind by computers? If science (and even more computer science) is the expression of our rationality, why are we, rational beings, caught off guard? Why at present we are not capable to clearly recognize the central problem of this crisis?

In my opinion, the happening of AI as a social problem was essentially a surprise because at present time mankind lacks self-consciousness. Indeed, all past experiences of useful reflection on the scientific enterprise failed:

> (1) The fight between the Church and Galileo Galilei detached science from the faith and hence the wisdom accumulated in previous millennia. Galilei supported the conception that "l'intenzione dello Spirito Santo essere d'insegnarci come si vadia al cielo, e non come vadia il cielo" (the intention of the Holy Spirit is to teach us how to go to heaven and not how the heaven goes). (Galilei 1978, pp. 128-135) Afterwards, the two worlds, the religious one and the scientific one, followed two distinct historical paths, without mutual communication. However, the latter world accumulated a so great number of novelties to lead to consider as largely surpassed the religious world with all its ancient sacred texts. Which therefore modern society has considered useless for orienting mankind's evolution.
>
> (2) Science born in ancient Greece in strong connection with philosophy. Yet, modern philosophy failed to understand contemporary science (its historical origin, the absolute notions of

Newton’s mechanics (space, time and gravitational force), its relationship with mathematics, and its foundations). Eventually Kant’s claimed to have offered a solution, which however promptly failed faced with chemistry and thermodynamics, and later it was ignored by both Einstein and thereafter the builders of the new physical theories. No surprise if scientists detached themselves from the traditional philosophy of knowledge by considering it a serious failure of modern Western philosophy.

(3) If modern science is the best representation of the rationality itself, why after its birth it did not progress in a linear and plainly cumulative way? In the early years of the 20th century science suffered a so radical crisis to lead a seemingly entire science’s failure. Why it suffered a crisis? However, the above-mentioned crisis was solved and moreover it was solved in such a brilliant way that e.g. quantum mechanics, notwithstanding concerning a new invisible, microscopic world, is a so perfect theory that it was the first physical theory without any divergence between its predictions and the experimental data. Why after this crisis science was even more powerful? No answer is known. Actually, some scholars attempted to understanding if not the nature of science, at least its history. Since the ‘50s Alexander Koyré, then Thomas Kuhn and Imre Lakatos offered historical appraisals of respectively the birth of modern science and the historical development of classical physics. But subsequent radical criticisms to their results led to evaluate them as insufficient attempts of inaugurating a new historiography, without having suggested any decisive improvement of this discipline[2].

[2] We have no answer even by applying Koyré’s main category (Koyré, 1957). The happening of AI is not characterized by the passage from finite to infinity which occurred five centuries ago and which produced a change that is commonly considered essentially positive and progressive. Moreover, by assuming Thomas Kuhn's interpretation of the history of science as composed by a series of scientific revolutions (Kuhn, 1969), the present revolution was not preceded by the discovery of any anomaly in the applications of previous paradigm.

Therefore, also the following question remains unanswered: Why the study of the history of past science has not led us to predict anything similar to AI's problem?

(4) Scientific research is developed in seemingly all directions; why did not it prevent us? Therefore, also the following question remains unanswered: why abruptly the absolute certainty communicated by science changed into a painful uncertainty on the survival of mankind itself?

(5) In particular, the philosophical reflection on computer science is still at a so low level that the question "Is computer science a science?" (Denning, 2005; Eden, 2007) does not receive a decisive answer. Moreover, no one knows why this theory is based instead of already performed and repeatable calculations in a finite span of time, on a modal word (computable functions) alluding to calculations which are considered as merely possible along an undetermined span of time. In addition, its theory is based on a proposition, Turing-Church thesis, which falsely is a thesis, because it precedes no theorem, for the plain reason that a theorem equating a formal notion (of computability) with the corresponding intuitive notion cannot exist.

(6) Modern technology, being informed by modern science, was so successful to radically transform the entire earth environment and hence the way of living of mankind so much to induce an anthropological change. Therefore, mankind's ways to tackle the reality are so radically changed that traditional ethics is de-valuated or even abandoned, without being replaced by any well formulated ethics (e.g. consider the "pill" and more in general human sexuality).

In sum, the dramatic question of AI occurs in a time of both an enormous human's capability of knowing scientific data and great expansion of the intellectual representation of the world, but at the same time of a lack of human awareness of his present historical situation. There could have been no worse time for mankind to deal with a so crucial question as AI.[3]

Some light comes only from a mere parallelism between the present process of interaction of humans with AI and the historical process of animal domestication. The advancements to AI oblige mankind to live together a multitude of computers capable to acquire an autonomous behavior in a similar way in past history it was obliged to live together with a multitude of bad and good animals. It was necessary a long process of more than ten thousand years for succeeding in domesticating some among the crowd of animal species. However, it was impossible to domesticate some species, called therefore "ferocious"; humans had to hardly fight for throwing out them from the human environment.

However, in this parallelism there exists also a difference in the possible peril of an unsuccessful "domestication process". In past time each person was the sole responsible of his project of domestication of an animal species; the peril (even of a death) consequent to its failure was individual one. Instead in the case of the AI the peril of an unsuccessful "domestication process" may have unforeseeable

[3] The great historical attempt of building a political alternative to Western capitalist society, i.e. the socialist movement, shared with capitalism the common faith in the scientific progress. Hence, no important attempt to criticize current science came from it. Only some thinkers launched alarms. The great sociologist Max Weber feared that the rationalization of society could lead to close the person "inside an iron cage" (Weber, 1930: 181). Maybe the most famous alarm was Martin Heidegger's desperate exclamation: "Only a god can save us!" (Heidegger, 1978). But no one suggested a way out these perils. Outside the West, Mohandas Gandhi influenced many in the world about the perils of machines and technology but did not produce an alternative attitude towards modern science.

consequences even on both the entire society and the next generations of humans, till up to the suppression of human species. Hence the present challenge is much higher than the past historical process of domestication of animal species because today the perils of even a single robot may concern the entire mankind.

In the following paper, I will show a limit to the activity of AI. In sect. 2 I suggest a clarification of a notion which is a preliminary to present research: "intelligence"; here it is referred to, rather single acts, the logical building systematic theories; eight degrees of logical intelligence result. In sect. 3 I suggest that among them the fourth one is the most relevant to the present investigation: to recognize the doubly negated propositions (DPNs). I quote as examples Isaac Asimov's principles of robotics which all are DNPs. Moreover, I offer the result of my long practice of recognizing DNPs within scientific and literary texts: a typology of all kinds of DNPs. In sect. 4 I will prove that a robot is essentially unable to recognize a DNP, owing to the incommensurability, marked by the failure of the double negation law, between its classical logic and intuitionist logic. In sect. 5 I show that a DNP implies a new kind of systematic organization of a theory which is aimed at discovering a new scientific method for solving a given problem. In sect. 6 the application of all in the above to AI proves its essential incapacity to implement any element of the new way of organizing a scientific theory Therefore I suggest assuming as a theoretical principle this "impotence" of AI.

2. Preliminaries: Definitions of both Meaning and Intelligence

The debate on AI is difficult because often the viewpoint of a discussant is not sufficiently manifested. It may be that of an individualist person, of a consumer, of a seller, of a jurist, of a political leader, of a spiritual leader, etc. Actually, the social dimension of the problems put by AI is essentially at the world level and also at the level of the future generations. For the first time in 20 Century the social advancements of mankind imposed the viewpoint of the entire mankind as a species and also as the best species. This viewpoint requires to each human a greater attention to the world life and its innumerable problems; therefore the process of growth of mankind consciousness at this level is a difficult process and at present it is not completed. No surprise then if the debate about the question of AI this viewpoint is not considered by many scholars of AI. In the following I will take into account only the human species viewpoint as the more appropriate one for adequately discussing the problem in the more general terms.

Let us now tackle a great problem, the definition of "meaning". A possible definition is the following one: "The meaning is the relevance to the action and thoughts that human attribute to the stimuli that they encounter in sensations". Of course, this definition gives the highest importance to the environment, which may be both the external one and the internal one of a person. Therefore the main problem for giving meaning to a proposition is to contextualize it. Whereas in order to contextualize a proposition humans make use of prior knowledge, a computer has to learn what in each case may be the context of reference. At first sight it seems impossible to enumerate all such contexts and submit them to machine learning.[4]

[4] The book (Landgrebe and Smith, 2022) suggests convincing arguments about this point. But is insufficient about AI potentialities because 1) it is based on mathematical arguments, yet mathematics can evolve in ever new theories; 2) it is based on the mathematical models of a complex system, whose theory however is not complete and stable; 3) more in general, it does not consider logical issues. Hence, it can suggest no more than plausible results.

No less difficulty presents the problem of defining the crucial notion of "intelligence". As an example, I suggest an old definition: "Intelligence is an aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with its environment". (Wechsler. 1944).

However, I suggest that a distinction exists between a person's single intellectual act, which may also result from a casual behavior and the construction of a systematic scientific theory. In correspondence to this distinction one can advance two main theses about AI: the *weak thesis*, according to which AI can have a little bit of human intelligence through some acts, and the *strong thesis*, according to which AI encompasses the entire human intelligence.

Under this light I advance the proposal of the following eight degrees of human intelligence ranging from the mere intelligence of single acts to the capability of human mind to construct systematic theories.

1st: To say a sensible proposition.

2nd: To respond in kind.

3rd: To reason by connecting propositions in classical logic (deductions, proofs of theorems).

4th: To formulate an axiomatic theory.

5th: To distinguish the two different types of logic, classical logic and intuitionist logic (to which the modal logic through its S4 model is equivalent (Hughes and Cresswell, 1996: 224 ff.)).

6th: To reason in at least two different kinds of logic, classical logic and intuitionist logic, in agreement with the two ways of reasoning suggested by both Plato (νοησισ and διανοια) and Nicolaus Cusanus (*ratio* and *intellectus*).

7th: To formulate an entire theory according to an inductive logic, the intuitionist one.

8th: To translate from one type of logic to another.

3. The Fifth Degree of Intelligence: Recognition of the Intuitionist Logic Within a Text

At present time AI is capable of the three first degrees of intelligence. It is disputable if AI can formulate an axiomatic theory (Robinson and Voronkov, 2001)

However, let us consider the fifth degree of intelligence: to *decide whether an author of a literary text is arguing in non-classical logic or not.* If one takes as a test an author's use of the law of the excluded middle, this decision is possible only when the author explicitly warns the reader of his use - or his rejection - of an exclusive or; surely, this case is rare in the literature. Instead, recent studies (Prawitz and Melmnaas, 1968); Grize, 1970; Prawitz, 1976; Dummett, 1977: 24) suggest that the best way of discriminating between the two kinds of logic is to refer to the failure of the double negation law; this failure introduces to argue within a non-classical logic; the most relevant one is the intuitionist logic. In addition, the difference of the two kinds of logic is not surmountable neither approximal (Gödel, 1933f).

Therefore, we are led *to take as a test an author's use of the double negation law.* As a fact the recognition of this logical figure inside a text is a not difficult task; one has to recognize following four aspects of the text: (*i*) the negative words, which often are apparent; (*ii*) then the propositions including two negative words; (*iii*) the propositions of this kind which are not equivalent to the corresponding affirmative propositions because the latter ones are lacking of supporting evidence (DNP); (*iv*) the essential role played within the text by each of the resulting propositions.

Hence, the crucial step of this method is to decide whether a doubly negated proposition is a genuine DNP. This task requires an accurate scrutiny because a variety of linguistic forms of doubly negated propositions exist. As a result of my long experience of investigating the occurrences of DNPs inside several kinds of texts I suggest an accurate and general method for inspecting DNPs inside a text. First of all, one has to recognize three cases of doubly negated propositions which are not DNPs and hence have to be discarded:

> (a) a doubly negated proposition which is a merely rhetoric one, because a verification of its correspondence with reality makes apparent that this proposition deals with an objective fact; in such a case the two negations do affirm, according to the classical law of double negation; hence, the proposition belongs to classical logic; e.g. "I have nothing else five euro"; "This move does not lead to you out of the room".
>
> (b) A doubly negated proposition which is equivalent to a merely negated proposition because one negation is reinforced in a psychological way by the other one; e.g. "I cannot go no further"; "I cannot get no satisfaction".
>
> (c) A doubly negated proposition where a negation explains the other one, e.g. "I have no answer from this deaf (wo)man" (actually this proposition is a shortened version of two propositions: "I have no answer from this (wo)man because (s)he is deaf"). Hence this proposition includes only one negation, the first one.

Thereafter, I suggest the following typology of all forms of a DNP; it has considered also the doubly negated propositions which are obscurely or ambiguously formulated; they are translated into accurate DNPs. Since the work of deciding the negative nature of a word is the first step of this inspection of a text, the following typology is ordered according to the increasing degree of attention required for defining a negative word.

A doubly negated proposition is a genuine DNP when it includes:

(*i*) Two independent negations; e.g. “It is not true that is not …”, or two independent negative words, e.g., “Unreal proposition are rejected”; “without contradiction”.

(*ii*) A single word composed by two negative words; e.g. "un-moving" (≠ fixed) and "in-variant" (≠ constant).

(*iii*) A word synthesising a DNP (I will underline point wise this kind of word); e.g. "only", which means "nothing but…".

(*iv*) A negative word plus a modal word, which is always equivalent to a DNP; e.g. "possible = it is not true that is not…” Notice that when an author establishes “equality except for the following constraints…” that amounts to state that the equality holds true according to a modality and hence we have not equality but a DNP of equality. Furthermore, notice that the common use of the words “equivalent” and “similar” covers a multitude of modalities of almost-equalities, without explaining in the case at issue which exact modality of equality is chosen.[5]

(*v*) A negative word plus an inequality – usually represented by a comparative adjective; the latter one is a negative word because it is

[5] Many authors of scientific theories wrote a proposition including a modal word plus two negations: e.g. “It is impossible a motion without an end.” This association of a modality with a double negation gives evidence for an author’s intuitive perception of the link between modal logic and intuitionist logic; this link has been formally established in recent times.(van Dalen, 1986: 300)

the negation of the desired case of equality. E.g. “more than…” and “less than…”. Notice that "it is not lesser than…" ≠ "it is greater than or equal to…" which is only one negation.

(*vi*) A negative word plus a question mark apparently waiting a negative answer; e.g. “Am I stupid? [Of course: No, I am not stupid]”; “Why not ?” [= I do not see a contrary reason]”.

(*vii*) A negative word plus an idealistic word which has cancelled (often in a euphemistic way) one more negative word; hence the latter one has to be restored; e.g. “create = produce from nothing”, “chimerical = not real”, “Platonic love = love in separation”, “perpetual = without an end”, etc..

(*viii*) A negative word plus an implicit negative one; hence, the latter one has to be discovered; e.g., "the [excessive] extent of this work perhaps hindered my countrymen…" (Lobachevsky 1840, Preface).

ix) A word conceived by the author as a negative one inside a given specific context; e.g. in theoretical physics the words “change”, “variation”, etc. are negative since they require theoretical explanations; in a mathematics context even the adjective “positive” may have a negative meaning, e.g. when it means a value “different from zero”. These examples stress that we have to overcome a main imprecision concerning the nature (either affirmative or negative) of a word which is a candidate for being recognized as negative word.

In linguistic it is usual to take into account more distinctions than the above ones. E.g., it is well know that there is a semantic difference between for instance “He is not a dishonest person” and “He is a not dishonest person”; “A not happy marriage” and “An un-happy marriage”. Furthermore the semantic difference between a negation of the whole predicate and a negation of the predicate term is underlined; e.g. “Not all males are un-married” and “All males are not un-married”.

Although semantically different, each couple of propositions includes two negative words; this is the logical fact which is noticed in an objective way by a reader; the remaining meanings of the propositions pertain to other worlds than the logical feature we are interested to. Therefore, I assume that the additional linguistic distinctions are too sophisticated in comparison to the basic distinction between the case of one negation and two negations; i.e. a proposition including two negations has to be counted as a DNP irrespectively of their locations inside the proposition.

To my experience the above cases exhaust the set of all cases one meets when one wants recognize inside a text DNPs. The relevance of DNPs to the subject of AI is shown by the following fact. All known principles of the ethics suggested for robots make use of this kind of proposition. Take as an example the celebrated Asimov's Principles; they all are DNPs:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Law 0. A robot cannot harm humanity, nor can it allow humanity to be harmed by its inaction.

4. AI Cannot Recognize the DNPs

Let us now consider the above listed three kinds of doubly negated propositions which are not DPNs. The reasons for discarding them as DNPs are of a semantic nature. In the first case the relation with reality plays a crucial role. AI cannot have cognition of (human) reality, but only with recounts of reality, with all the limitations and deformations implied by recounts; moreover, the reality is too much large to be entirely assumed by a computer's memory.[6]

In the second case the proposition at issue includes a psychological emphasis whereas AI cannot intend this emphasis; it only sees the two negations. In the third case the latter negation does not counts as a negation because it merely explains the nature of the former one; to recognize their linkage as an explanation requires to understand the meaning of the former negation; that is impossible to AI.[7] *As a first result we obtain that AI cannot distinguish the genuine DNPs from the false ones*. As a consequence, any human discourse based on doubly negated propositions may be reiterated by AI, but without sharing its semantics; AI constitutes no more than an extremely advanced parrot.

Moreover, notice that in the cases *i)-v)* one recognizes the DNPs by applying syntactical rules. Instead in the cases *vi)-ix)* one recognizes the DNPs by understanding the meanings of the propositions. Among the latter ones the last cases *viii)-ix)* are the most difficult ones because the recognition of the nature of a DNP depends from the semantics of the

[6] Recall that a computer translating English-French needs 36 millions of couples of propositions.

[7] "It seems that language acquisition presupposes the working of a common set of ontological distinctions on the side of language learners, including the distinction between object and process, between individuals and categories, between natural and accidental properties of objects, and so forth." (Landgrebe and Smith, 2021: 206)

context of the proposition under examination; hence, one has to pay attention to an a priori undefined number of other propositions of the text. As a consequence, an accurate analysis of all the occurrences of DNPs inside a text requires to a human to reiterate its reading in order to be eventually sure to have grasped all the possible meanings of both the words in the propositions and the context of them.

Let us apply these remarks to AI: First, AI cannot recognize the false DNPs of the three cases listed in the above. Moreover, since the variety of genuine DPNs is infinite it cannot be implemented by a computer. In addition, whereas the DNPs *i*) – *v*) are easily implemented by AI thorough their syntactical rules governing them, in the remaining DNPs *vi*) – *ix*) AI does not perceive the negative nature of at least one negative word, because this nature is essentially derived by semantic criteria.

One may decide to ignore this deficiency and proceed with AI managing only the DNPs *i*) – *v*). In such a case we have a computer whose behavior is closed inside a *schizophrenic isolation*. It may be useful for some restricted tasks, but when located within an open (human) context its results are unrealizable.

As a consequence, Turing test is overcome. It is enough to submit to a robot a doubly negated proposition as a question "The <u>enemies</u> of your <u>enemies</u> are enemies or friends?" The robot can produce a disquisition on the double negation and even suggest that according to the two kinds of logic, classical and intuitionist, there exist two opposed answers; yet, it cannot decide on the dichotomy because, depending the submitted proposition from the context (essentially war/peace), the question cannot be solved by it if not by applying classical logic, i.e. by answering "Friends", which is not always the correct answer.

Rather, I suggest a new test of intelligence which consists in recognizing a DNP. Moreover, since more than half a century computer programming was unable to implement a DNP, I suggest that the time has come to decide that this implementation is impossible and thus suggests a principle of the impossibility of a computer implementation of a DNP, like in the history of theoretical physics the Academie de France in 1775 decided to reject all proposals of new machines implementing a perpetual mobile.

5. A New Formal Organization of a Scientific Theory: Its Inaccessibility by AI

From a comparative analysis of some scientific theories which respective authors presented in a non-axiomatic way (e.g. classical Chemistry, Lazare Carnot's mechanics, Sadi Carnot's thermos-dynamics, Lobachevsky's non-Euclidean geometry, Einstein's first theory of quanta, Dirac's quantum mechanics, Kolmogorov's formalization of intuitionist logic), Drago (2012) extracted the model of an alternative theoretical organization to the deductive-axiomatic organization; it is called a problem-based organization. Within it the basic propositions are DNPs.

In each above theory the DPNs compose *ad absurdum* proofs, whose final conclusion is again a DNP. To it the principle of sufficient reason (PSR; Leibniz, 1686) is then applied, under two constraints (Markov, 1961: 5), in order to translate this conclusion into the corresponding affirmative proposition, the only one which can be tested with reality. This is the last step of such an organization of a theory.

Let us now consider the *ad absurdum* proofs. They argue by means of implications between DNPs. Hence, the arguing is not assuredly recognized by AI, owing to the insufficiency of AI in recognizing DNPs. Moreover, the conclusion of this kind of proof is essentially a

DNP; otherwise, it could be reverted into a direct proof (Grize, 1991). This component of this theoretical organization is one more insufficiently recognized feature by AI.

However, granted that the DNPs are well-recognized by AI, the conclusion is a DNP on which the PSR has to be applied under Markov's two constraints, in order to obtain an affirmative proposition. The second Markov's constraint on the conclusion of the final *ad absurdum* proof, i.e. to be decidable, is not easily fulfilled, because one has to recognize whether the conclusion has to be logically decidable, or mathematically decidable, or operatively decidable. The first two cases are problems also for humans; in particular, to decide that a proposition is undecidable requires a proof which has to be invented. The third case is in general unsolvable by AI owing to its essentially partial knowledge of the physical world.

In sum, for several reasons the distance of AI from a PO theory appears as insuperable.

6. Conclusions

In the above we obtained relevant answers to some introductory questions because we put remedies to the insufficiencies of our awareness of the subjects presented by the remarks 3) – 5), i.e. the history and the foundations of science plus the philosophical reflection on the limitations of a computer.

In general terms, two millennia of scientific theories showed that there only exists two kinds of theoretical organizations and moreover last century research showed that they are severed by the insurmountable difference between classical logic and intuitionist logic, i.e. the validity or not of the law of the double negation. We have to conclude that computer have no capabilities to represent both a problem-based organization and intuitionist logic. In the history of scientific theories this half part of human intelligence was very productive (thermosdynamics, non-Euclidean geometry, the revolution of theoretical physics in early ‘900, etc.); actually, its surprising results show that it represents the highest power of human intelligence.

Since human mind can choose the kind of logic and in particular may follow intuitionist logic according to a problem-based theoretical organization, its capabilities cannot be overcome by a computer and AI either.

In conclusion, all in the above proves that the common way of conceiving as valid only classical logic, merely because it appears as the most productive one explains in a great part the AI problem. AI seems equate the performances of human intelligence because it maybe equates the human reasoning in classical logic, but human mind includes an alternative logical world. Hence, the introduction of AI leads humans to depreciate the arguing in classical logic which AI may manage and implement, for rather mainly develop the arguing in intuitionist logic. Moreover, it suggests a new principle, that of an impotence of AI.

I add a final consideration of historic-philosophical nature. The traditional attitude for an absolute certainty of classical logic and its indispensability obscured the recognition of the alternative ways of both reasoning and organizing a theory. In the past time only few scholars have explored the new computing machines. One of them was Charles Peirce who stressed “computer’s impotencies”. At present we see that the real content of Peirce’s appraisal is that AI cannot represent a problem-based organization of a theory and also its logical components. That makes clear the “impotencies” of the “intelligent machines” that Peirce recognized in anticipation.

References

Denning, P.J. (2005). "Is Computer Science Science?". *Comm. ACM*, 48, no. 4, pp. 27-31.

Drago, A. (2012). "Pluralism in Logic: The Square of Opposition, Leibniz' Principle of Sufficient Reason and Markov's principle". In: Béziau, J.-Y. and Jacquette, D. (eds.). *Around and Beyond the Square of Opposition.* Basel: Birkhaueser, pp. 175-189.

Dummett, M. (1977). *Principles of Intuitionism.* Oxford: Clarendon Press.

Eden, A.H. (2007). "Three paradigms of Computer Science". *Minds and Machines*, 17, pp. 135-167.

Galilei, G. (1978). *Lettere.* Torino: Einaudi.

Gardiès, J.-L. (1991). *Le raisonnement par l'absurde.* Paris: PUF.

Goedel, K., (1933). *Collected Works.* Oxford: Oxford University Press., 1986, papers 1932 and 1933f.

Grize, J.B. (1970). « Logique ». In : Piaget, J. (ed.). *Logique et connaissance scientifique.* Éncyclopédie de la Pléiade. Paris: Gallimard, pp. 135-288.

Heidegger, M. (1976). "Nur noch ein Gott kann uns retten". *Der Spiegel*, May 31st.

Hughes, G.E. and Cresswell, M.J. (1996). *A New Introduction to Modal Logic.* London: Routledge.

Koyré A. (1957), *From the Closed Cosmos to the Infinite World*. Baltimore: Wisconsin Univ. Press.

Kuhn T.S. (1969), *The Structure of the Scientific Revolutions*. Chicago: Chicago University Press.

Landgrebe J. and Smith B. (2021), "Making AI meaningful again", *Synthese*, 198, pp. 2061-2081.

Landgrebe J. and Smith B. (2022), *Why Machines Will Never Rule the World*. London: Routledge

Leibniz, G. W. (1786). *Letter to Arnauld.* 14-7-1686, Gerh. II, Q., Opusc., pp. 402, 513.

Lobachevsky, N.I. (1955). *Geometrische Untersuchungen zur der Theorien der Parallellineen.* Berlin: Finkl, 1840 (English transl. as an Appendix to Bonola, R., *Non-Euclidean Geometry.* New York: Dover).

Markov, A.A. (1962). "On Constructive Mathematics". *Trudy Math. Inst. Steklov*, 67 8-14; also in *Am. Math. Soc. Translations,* 98 (2), pp. 1-9.

Peirce, S.C. (1931-1935). *Collected Papers of Charles Sanders Peirce* [*1931 — 1935*], Cambridge: Cambridge University Press.

Prawitz, D. (1976). "Meaning and Proof. The Conflict between Classical and Intuitionist Logic". *Theoria*, 43, pp. 6-39.

Prawitz, D. and Melmnaas, P.-E. (1968). "A survey of some connections between classical intuitionistic and minimal logic". In: Schmidt, A. Schuette, H., Thiele, E. J. (eds.). *Contributions to Mathematical Logic.* Amsterdam: North-Holland, pp. 215-229.

Robinson, A. and Voronkov, A. (2001). *Handbook of Automated Reasoning.* Cambridge MA: MIT Press.

Russell, B. (1903). *The Principles of Mathematics.* Cambridge: Cambridge University Press.

Weber, M. (1930). *The Protestant Ethic and the Spirit of Capitalism*. New York: Routledge.