

Antonino Drago¹

Full professor

Faculty of Philosophy

University Federico II of Naples

ETHICS OF ROBOTS

Abstract: *This paper explores a multidimensional framework rooted in two foundational dichotomies: the distinction between potential infinity and actual infinity, and the organizational structure of scientific theories—either as deductive systems from axioms or as problem-oriented approaches. These dichotomies serve as a theoretical lens through which various domains—science, ethics, biology, and technology—are analyzed. Beginning with a critique of modern set theory and its reliance on actual infinity, the work contrasts it with the constructivist stance grounded in potential infinity, arguing that each implies different epistemological commitments and logical structures, particularly regarding undecidability and the law of double negation. The paper culminates in a critical reflection on the ethics of human-robot interaction.*

Key words: *Scientific Theory Models, Intuitionist Logic, Human-Robot Ethics, Consequentialist Ethics, Potential and Actual Infinity.*

¹ drago@unina.it; ORCID: 0000-0001-6002-1258

1. The basic assumptions

A) It is well-known the ancient dichotomy on the infinity: either potential infinity, PI, or actual infinity, AI. The recent result of the infinity of degrees of powers which are intermediate between the discrete and the continuum was obtained by idealistic means (Cohen 1964), hence it is valid by assuming just AI. The two kind of infinity are mutually exclusive because by assuming PI, some idealistic hypotheses (e.g. Zermelo's axiom) are unreachable by finite means. *Viceversa*, by assuming AI each undecidability is solvable by an idealistic hypothesis, provided that it is not contradictory with the previous framework.

B) *A dichotomy on the kind of organization of a scientific theory.* The study of some past scientific theories showed that some theories have been built by their authors on a different model of organization from that of the deductive theories. the common model of all these theories was characterized by the following steps of their logical developments:

i) A problem which is unsolvable through usual tools is declared; e.g., in Lobachevsky's geometry: how many parallel lines exist.

ii) No more than the common knowledge on the field at issue is presupposed.

iii) The theory is aimed at discovering a new scientific method capable to solve the given problem.

iv) The theory makes use of doubly negated propositions whose corresponding affirmative propositions are not equivalent in meanings (DNPs); hence, double negation law fails and they belong to intuitionist logic.

v) By composing together DNPs the theory argues through *ad absurdum* arguments (AAA); the conclusion of each AAA is again a DNP (i.e. here an AAA is a “weak” *ad absurdum* proof); it may work as a premise for a next AAA, so to obtain a chain of AAAs.

vi) The conclusion of the final AAA is a universal predicate, $\neg\neg UP$, which represents a possible hint for the resolution of the given problem and all related problems.

vii) At this point as a matter of fact the author translates the above predicate $\neg\neg UP$ into the corresponding affirmative predicate UP . Apparently the author thinks to have collected by his previous reasoning enough evidence to be justified in promoting his conclusion $\neg\neg UT$ to the corresponding affirmative proposition UP , although this change is not allowed by intuitionist logic, which previously he had adhered to.

viii) Exactly because now this proposition UP is an affirmative one, the author can test against reality it and all the derived propositions derived from it (by means of classical logic); hence, he can validate his entire theory or not (Drago 2012).

In conclusion, there exist two kinds of theoretical organization (either the organization aimed at solving a basic problem PO, or a deductive organization from few principles-axioms AO). Owing to the failure of the double negation law they constitute one more dichotomy.

C) *Four models of scientific theory.* Since the year 1988 I presented the above two basic dichotomies of the scientific theories (Drago 1988). These dichotomies constitute the foundations of all physical theories. The couples of choices on them fashion four models of

scientific theory, whose representative physical theories are Geometrical optics (PI&AO) Newton's mechanics (AI&AO), Thermodynamics (PI&PO) and Statistical mechanics (AI&PO). These models are incompatible (and the representative theories are mutually incommensurable) owing to the choices of each dichotomy are mutually exclusive.

D) *Four models of Ethics*. Some decades ago Elio Sgreccia recognized four models of ethics as those followed by people in the world (Sgreccia 1986). They may be traced back to two dichotomies, whose contents correspond to those of the scientific theories; one dichotomy concerns the notion of an infinite aim (either a potentially achievable aim or an actually infinite, mythical aim) and the other dichotomy concerns the notion of organization (either aimed at solving a basic problem or as a deductive system from few ethical principles). The four couples of choices fashion four models of ethics which are mutually incompatible owing to their different basic choices. That means an irreducible pluralism of ethics; however the followers of these different models of ethics may converge on establishing some common norms. The correspondence between the ethical choices and the scientific choices is so tight to be denoted by the same acronyms, PI and AI, PO and AO.

I trace back this pluralism to Leibniz's suggestion of two human mind's labyrinths: actual infinity or potential infinity; law or freedom; the latter labyrinth is interpreted as a subjective way to make experience of organizing a multitude of elements into an entire system. Since he was unable to decide them by rational tools and even along three centuries next scholars were unsuccessful, we are allowed to conclude that his labyrinths are two dichotomies.

E) *Definitions of an organism, a human being and an automaton.* A scholar wrote the following words: “The secret of life is the ability of living organisms to improve [indefinitely] their [organizational] programs” (Longuet-Higgins C. 1969, p. 227). Mayr’s characterization of life (Mayr 1996, pp. 101-102) includes the following characteristic features: “code”, “complexity and organization”, “order”, “non-constancy”, “evolution”. As a fact, the notion of infinity includes “code” and “evolution” (plus “non-constancy”); whereas the notion of organization summarizes “complexity” and “organization”, and “order”. By following a suggestion by Carl Sagan, the NASA committee proposed that life is a “[specific organization] self-sustaining chemical system capable of [actual infinity] Darwinian evolution.” (Benner 2010) In conclusion, the above definitions indicate the additional characteristics of a living organism with respect to an inanimate thing: 1) its own exclusive organization (metabolism, etc.); 2) its codified information that extends its life along an infinite span of time through its posterity projects and over an unbounded space.

Second, these dichotomies suggest a definition of a human with respect to an animal. Whereas an animal only experiences a potential infinity (PI) and a fixed organization which is aimed at solving a basic problem (PO), i.e. the survival, a human adds to the choices PI&PO constituting the fixed life pattern of an animal the opposite choices on the two dichotomies, i.e. AI and AO. They pertain to the culture since a human may imagine AI as a thing actually existing beyond any physical bound and AO as an ideal organization of an infinite pyramid of consequences (from few principles-axiom), although this pyramid cannot be grasped in its entirety. Surely,

imagination is not reality, but a human may rise up his thinking from reality to imagination (e.g. geometry) and then elaborate the imagined things so much to generate new hypotheses which can be tested with reality, i.e. at last the imagination may surmise new physical results. This capability of imagination gives to humans a powerful tool for changing the given reality.

Third, these dichotomies suggest a definition of a robot with respect to a human. A century of working in computer programming shows that a robot is unable to implementing a doubly negated proposition, and hence the entire PO. Moreover, it is unable to establish an undecidability (apart the theoretical use of idealistic operations as an oracle or an unbounded minimalization). Truly, Computer science was built on modal words (e.g. *computable* functions) and idealistic operations (as the two above mentioned ones), but a robot cannot appeal to this human theory for implementing its capabilities. Hence Turing test is fulfilled. In conclusion, the robots are essentially bounded with respect to a human capabilities, who may use without bounds of AI and AO (till up to trespassing into a metaphysical realm, at all detached from reality). Notice that already in 19 th century Peirce stressed the “incapacities” of the “intelligent machine” (Peirce 1887).

F) *A comprehensive ethics of the interaction's human-robot.* However, the superiority of a human over robot has no relevance with respect to the great perils generated by the introduction of robots in social life, because the bad intentions of a whatsoever human can be implemented by a robot into gigantic disasters. Hence the very problem of the present introduction of robot in social life is not the robots by themselves, but the interaction human-robot. Since the perils of the human-

robot interaction may be caused by whatsoever human, the ethics to be promoted has to be of a universal kind, i.e. compulsory for each man of the present mankind. Moreover, since the perils of ill directed robot may be of exceptional gravity, the ethics to be established has to introduce some absolutely negative commands for all men (taboos) and even (as in the case of nuclear weapons) for the next generations. This kind of ethics is new with respect to all previous ethics for spatial and temporal extension. The target may seem too ambitious, but (Jonas 1972) claimed that this kind of ethics is adequate to the new anthropological age inaugurated by the exceptional technological progress of the 20 century

G) *A pluralism of ethics can converge all together for facing all together a major peril?* Which is the maximum peril? It is the peril of the suicidal of mankind the main threat to mankind's survival. First of all, in a nuclear age this survival is clearly threatened by world wars. Let us recall that during a war each opponent applies all tools for performing killing and massacres, mainly in the perspective to be the loser. In such a case all sophisticated tools are employed in order to grow exponentially the massacres. Then, no ethics about robot makes sense if the humans persist in making war which in all cases may escalate to world war. If the new ethics has to avoid this perspective, no war has to occur on the earth, with penalty that ethics, aggravated by the presence of the robots, cannot be applied at all. Hence, no more the *jus in bello* is meaningful when disposing of ever more advanced killer-robots. The birth of United Nations leaves to hope that we can start this era where no longer a war will be elicited. The same holds true for the civil life within a society, which however is already ruled by a strict application of the jurisprudence about the killing within the interpersonal relationships. The program is

then to extend the specific laws of civil society to the international society of the anarchical interacting of states. In conclusion, the unique way to bound the perilous potentialities at disposal of a man equipped by computers, is to absolutely avoid any killing and any attempt to kill. Only after this crucial step one may study the interaction man-robot with a minimum hope to rule this new situation without fearing that one has to face at massacres. In other words, the new ethics has to start a new age in which mankind has to radically change some old, brutal customs.

H) Since its main attention is to avoid possible killing, a new ethics becomes very relevant; it is no longer an ethics based upon ideal values, or upon the specific situation of a set of norms, or a merely subjective viewpoint; it is a consequentialist ethics. In the 20 century an important consequentialist ethics is born, the ethics of non-violence (M.K. Gandhi); it is based on the choices PI (un unlimited number of inter-personal relationships) and PO i.e. it is based on the problem, how to look for cooperative ways to solve conflicts.

H) First of all, when one thinks a robot one has to choose which kind of theory of his interaction with it he wants to build. In my opinion no axiomatic theory is suitable for fulfilling our purpose. Rather the wanted theory has to be based on a basic problem, i.e. the problem that we recognize at glance: may the autonomous behavior of a robot lead to an acceptable behavior to human beings? A theory based on a problem is managed by not a deductive reasoning, but an inductive reasoning. As a consequence, we have to reason not in a deductive way, but in an inductive way. That means that the intuitionist logic governs the wanted theory; hence,

most its propositions are doubly negated ones. In fact, all known principle of the ethics applied to robot make use of this kind of proposition. Take as an example the celebrated Asimov's Principles; they all are doubly negated propositions:

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. (Asimov 1950)

I) Can one build a theory of this interaction by suggesting AAAs from the basic DNPs? And then apply the PSR to their conclusion? This goal is too ambitious at this first stage of research. Rather, let us proceed in a more pragmatic way by generalizing what is the thousand- year ethical heritage which has been assumed by all national legislations: the four social commandments. They want to rule all human inter-relationships; maybe they may be suitably extended to the new relationships of human with robot.

Surely, "*Thou shalt not kill*" has to be included in this ethics for the reasons illustrated in the above. It wants to prevent any disastrous War. "*Thou shalt not commit adultery*" may be intended as forbidding the possibility of a robot to present itself in a misleading way. It wants to prevent any Sedition or Revolution of robots. "*Thou shalt not steal*" may be intended as forbidding the robot's social acquisitions replacing humans from their natural places. It wants to prevent any fallen of humans into Misery. "*Thou shalt not bear false witness*" may be

intended as forbidding a cumbersome or even misleading inspection of the robot's program. It wants to prevent any Servitude of humans. (notice that the four social plagues listed in the above have been theorized by (Lanza del Vasto 1959, first page), and reduced to the expressions of two basic dichotomies.

One may hope that the completeness of previous commandments in preventing all possible negative actions within the human interactions is preserved by their above extensions to robot- human's interactions.

References

1. Benner S.A. (2010), "Defining Life", *Astrobiology*, 10, pp. 1021-2030.
2. Cohen P.J. (1963), "The independence of the Continuum Hypothesis, [part I]". *Proceedings of the National Academy of Sciences of the United States of America*. 50 (6): 1143– 1148.
3. Drago A. (1988), "A Characterization of Newtonian Paradigm, in P.B. Scheurer, O. Debrock (eds.): *Newton's Scientific and Philosophical Legacy*, Kluwer Acad. P., 1988, pp. 239-252.
4. Drago A. (2000), "Etica e scienza: loro fondazione comune secondo una visione pluralista", in L. Chieffi (ed.): *Bioetica*, Paravia, Scriptorium, Torino, 303-331.
5. Drago A. (2012), "The emergence of two dichotomys from Einstein's first paper on Quanta", in Pisano R., Capecchi D., Lukesova A. (eds.), *Physics, Astronomy and Engineering. Critical Problems in the History of*

- Science and Society*, Siauliai: Scientia Socialis P., 2013, 227-234.
6. Jonas H. (1979), *Das Prinzip Verantwortung*, (Engl. Tr. *The Imperative of Responsibility*, Chicago: U. Chicago P., 1984.
 7. Lanza del Vasto, *Les Quatre Fléaux*, Denoël, Paris, 1959.
 8. Longuet-Higgins C. (1969), "What biology is about" In Waddington C.H. (ed.) *Towards a Theoretical Biology*, vol. 2, p. 227. Edinburgh: Edinburgh University Press.
 9. Mayr E. (1972), *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*, Lakonia NH: Belknap Press.
 10. Mayr E. (1996), "The Position of Biology among the Sciences", *Quart. Rev. of Biology*, 71, pp. 97-106.
 11. Peirce C.S. (1887), "Logical Machines", *Am. J. Psychology*, 1, 165-170, pp. 168-169.
 12. Sgreccia E.: *Manuale di Bioetica*, Milano: Vita e Pensiero, 1988, pp. 74-90. (Engl. Transl. *Personalist Bioethics. Foundations and applications*. Broomall PA: National Catholic Bioethics Center, pp. 43-53)