Bridging the Gap of Neuroscience, Philosophy, and Evolutionary Biology to Propose an Approach to Machine Learning of Human-Like Ethics

Nicolas Lori(✉), Diana Ferreira, Victor Alves, and José Machado

Centre Algoritmi, University of Minho, 4710-057 Braga, Portugal
{nicolas.lori,diana.ferreira}@algoritmi.uminho.pt
{valves,jmac}@di.uminho.pt

Abstract. The growing explosion of ideas such as Artificial Intelligence (AI), smart environments and ubiquitous computing has led to the creation of the Ambient Intelligence (AmI) paradigm. As AmI begins to take place, moving from a futuristic idea to a reality, we are gradually witnessing the creation of an omnipresent, responsive, and intelligent atmosphere in which thousands of tiny sensors and natural user interfaces will be embedded in our natural movements and in our social and physical interactions. Hence, a key challenge in this multi-disciplinary approach is to get machines to act according to ethical priorities that make sense to human beings. In this study, we improve the capacity for machine ethics to approach human ethics by assessing the computation of transaction values and we argue that it is possible to perform such a computation using recent work that describes the effects of human decision-making using an axiomatic framework. This paper clarifies the relationship between the brain's 3-axes of neuroscience, the 3 Plato's Transcendentals of philosophy and the biological evolution's 3-components, as well as the top-down vs. bottom-up approaches to machine ethics.

Keywords: Artificial Intelligence · Ambient Intelligence · Machine ethics · Transaction value · Aesthetics evolution · Plato's Transcendentals · Axiomatic systems

1 Introduction

Embodied by the combination of autonomous systems, AI, and information technology, the 4th industrial revolution has been promoting a permanent transformation of morals, knowledge, and perceptions in almost all areas of human expertise [16,22,23]. The ethical, economic, and social implications of this revolution are a worldwide concern and a matter of political and public deliberation [6,39], which are causing a reappreciation of how to compute the transaction value of an entity.

C. Analide et al. (Eds.): IDEAL 2020, LNCS 12490, pp. 309–321, 2020.
https://doi.org/10.1007/978-3-030-62365-4_30

A key aspect of [20] is the intricate relation between philosophy and computer science, and it was there proposed that the construction of such relation is greatly improved by the use of contemporary neuroscience. Based in the work of Schiller, the utmost Beauty-value should be assigned to objects that present the uppermost freedom [20], i.e., the objects that have the least usefulness and are therefore closer to a thing that exists for itself. In an opposing view, in [7] it is proposed that value is "useful information", translated as Knowledge and that is designated as Truth in [20]. Therefore, contemporary economics seems to place more value on Truth than Beauty, and people are prepared to spend their wealth to improve their bodily self-perception, which [20] associated to the concept of Good. Hence, there are three types of value: Truth-value, Good-value, and Beauty-value. Whereas, a recent book, "The Square and the Tower" [15], uses the theory of computer-networks to analyze the relation between history and contemporary socio-economics; where "Tower" means hierarchical command-and-control structures hence maximizing Truth-value, whereas "Square" symbolizes equalitarian distributed-control networks hence maximizing Good-value.

Ethical options are based on axiomatic choices promoted by cultures that can be either explicitly or implicitly religious, as they always require certain axioms to be valid without the support of an experimental validation. Thus, any approach to ethics is always about re-connecting (*religio* in Latin means "bind back") the mundane conditioned existence to a transcendental unconditioned valuation system that separates ethical from unethical. As noted, the balance between Truth-value and Good-value is a key characterizer of cultures [15,19,20]. In this work, the Transcendent is a generic term to signify God from the Christian perspective, or Absolute from the perspective of contemporary science.

Ethical choice perspectives are very important because they decide something that cannot be trivially decided by a computational binary logic of "Valid/1 vs. Invalid/0". Wittgenstein was the first to detail this fundamental gap of all non-religious thought at the end of his Tractatus Logico-Philosophicus [41]. However, despite Wittgenstein being the first, the Scottish Enlightenment of Adam Smith, Edmund Burke, and David Hume had already pointed out that the Enlightenment could only focus on data that could be made objective through detailed observation, thus leaving out religion and the arts that should be allowed to evolve over time. On the other hand, the Continental Enlightenment of Spinoza, Voltaire, and Schopenhauer argued that religion was obsolete and should in the future be replaced by art, with music being the most sublime aspect of art.

Until recently, it was thought that causalities were indeterminable and that all that could be statistically determined were correlations, a form of thinking that is well matched with Hume's perspective on ethics, but Pearl [27] obtained that it is possible to determine the direction of causality, and moreover, without determining it, the statistical analysis of events is necessarily wrong, and this determination of the need to define the direction of causality is more in line with Kant's ethical perspective, which is strongly supported by cause-effect relationships [19]. Unfortunately, this analysis of causality's direction is always difficult, and it is always probabilistic as it is based on Bayesian correlations. Aristotle's

final cause can be understood as Darwinian causality, which is directly related to random walking processes, and randomness was a topic that also interested Aristotle, who was very important for the Scholastics. Therefore, one would expect the Scholastics to have explored the statistical perspective of causality, but that did not happen. For Scholastics, the formal logical deduction was the unquestionable basis of all causality, and consequently, Christian thought never developed a statistical perspective of ethics. Moreover, Enlightenment thinkers sought a determinism based on Reason and so always considered randomness as a temporary indication of contemporary limits of observation, and never as something fundamental, and therefore did not develop a statistical perspective of ethics either. Nevertheless, there were in the ancient past references to ethics, occurring in non-deterministic ways (e.g. Tower of Siloam parable).

Enlightenment believed that human rationality could perceive all the causality of the universe because all determinism is determinable, it therefore forgot to consider the possibilities that: causality was not deterministic, that what we do not know the cause of had sometimes no cause, and that the non-contradictory may never be able to be universal. Unfortunately, for the Enlightenment, these three things happened, respectively being: Darwin's evolution (in 1859), Heisenberg's uncertainty (in 1927), and Gödel's incompleteness (in 1931). In this work, a path is followed that is radically different from the Scholastics and the Enlightenment, preferring the assumption of a stochastic nature of the universe, instead of a formal logic approach. This choice is made because, since the works of Darwin, Heisenberg, and Gödel, it makes no sense to use formal logical deduction as a basis. Therefore, it is proposed here that with the use of Philosophy of Information [17] it is possible to put statistics as the basis for the foundations of ethics. Hence, as computer iterations follow a formal logic inference [9], a simple axiomatic approach to computer ethics will have the same limitations as the Scholastic and Enlightenment ethics. Thus, this statistical ethics approach has direct applications to how machine ethics is developed and applied.

We will use a 3-axes value approach [20] to establish a relation between biological evolution, brain axes, philosophy, psychology, and axiomatic systems that will make it easier to develop more human-like machine ethics. In order to assess the 3-axes value, we must work with an "axiom-driven value calculation" for Truth-value, e.g. a "deterministic inference" Newtonian Axiomatic System [9,21]; and with an "environmentally-driven value calculation" for Good-value, e.g. a "natural selection" Darwinian Axiomatic System [21]; whereas a new form of partial information-driven calculation, is required for the Beauty-value, just like in biology the "aesthetic evolution" [28,29] constitutes a suggestion of a "best guess" representation of evolution's undeterminable [34] future. A relevant aspect for this new form of partial information-driven calculation called Statistical Philosophy (e.g. by using Shannon Information), a Philosophy of Information [20] branch, is the important role of causality in determining the appropriate statistical approach for the events [27].

These 3-axes, Truth+Beauty+Good [19,20] are what is known in philosophy as Plato's Transcendentals, and are equivalent to three major branches of philosophy through the relation: i. Epistemology searches for what is Truth; ii. Ethics searches for what is Good; iii. Aesthetics searches for what is Beauty. There are two other branches of philosophy: logic and metaphysics, but logic is simply the implementation of the equivalences to the axioms assumed as Truth by the epistemology, whereas metaphysics by definition aims at understanding what is beyond what physics can provide, and presently it is known that Darwinism can be considered to be beyond determinism [34], even beyond stochastic average determinism, and hence beyond physics. Thus, both philosophy and contemporary neuroscience perspectives can be reduced to a 3-axes system (see Fig. 1, and Table 1).

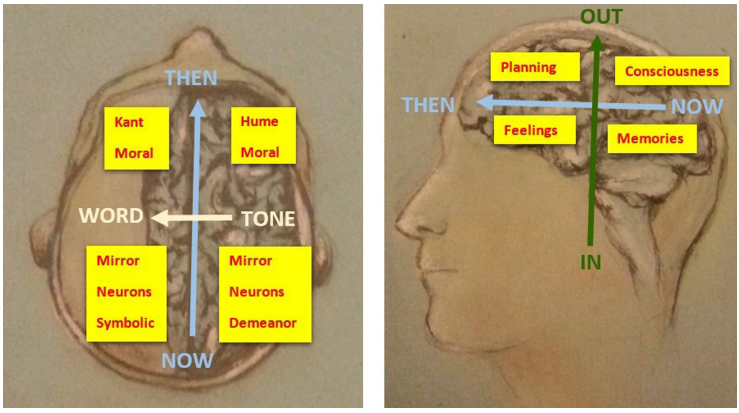

Fig. 1. Relation between brain 3-axes approach and brain function. On the left is a view from above, and on the right a view from the left side (®Sandra Lori) [20].

The simultaneous maximization of Truth-value, Good-value, and Beauty-value is the obvious goal; but, just as it seems impossible to have equipment costs, product improvement and wages maximized without going bankrupt; the simultaneous maximization of Truth-value, Good-value and Beauty-value often leads to a trilemma restriction that occurs in many forms, e.g. the Political Trilemma (triplet of "Democracy vs. national sovereignty vs. global economic integration") [30] and the Impossible Trinity (triplet of "independent monetary

Table 1. Plato's Transcendentals relation to contemporary neuroscience axes [19,20].

Transcendentals	Axes
Truth	Hunting/Power/Now-Then
Beauty	Choosing/Meaning/Tone-Word
Good	Eating/Pleasure/In-Out

policy vs. fixed exchange rate vs. free movement of capital") [8,26]. Finding the maximization balance in entity transaction management is, hence, the finding of an appropriate balance in the Political Trilemma and the Impossible Trinity.

2 Related Work

The gradual shift towards ubiquitous computing and AmI is responsible for the foundation of an omnipresent and intelligent atmosphere. As computers' decision-making roles grow and autonomous machines become more sophisticated, society increasingly relies on computer-based intelligence with reduced human supervision. Unquestionably, granting control and autonomy to machines requires them to act in an ethical way. Ethics is necessary to determine what is morally right or wrong [37], to be a factor in the attribution of responsibility [35], to decrease the likelihood of negative outcomes for humans and/or to narrow the adverse effects machines can cause. Hence, the growing demand to regulate intelligent systems and bring forth better ethical approaches. Machine Ethics seeks to implement a moral dimension in computational systems either by introducing moral principles in machines or by discovering means to make machines function in an ethically responsible way on their own.

As machine ethics is a combination of computer science and moral philosophy, the scientific literature includes publications of different natures, ranging from theoretical papers about what a machine can or should do [11,35], to experiments about the incorporation of ethical reasoning in computer systems [3,40]. Allen *et al.* identified a high-level classification to machine ethics based on the nature of the approach: top-down approaches, bottom-up approaches, and a hybrid of top-down and bottom-up approaches. A top-down approach requires earlier specific moral principles or theories to train the machine to identify ethically appropriate actions as well as to recognize and correctly react to ethical scenarios and dilemmas [2,38]. In contrast, a bottom-up approach does not impose specific moral principles or theories, instead considers moral values as being implicit in the activity of agents and tries to provide agents the power to understand their own morality and the morality of others [2,38].

There is a multitude of works dedicated to top-down approaches, such as Dennis *et al.* development of ETHAN, a system that deals with situations where civil air navigation regulations conflict with each other [12]. The system is provided with a particular ethical policy that refers to four hierarchical ordered moral principles (do not harm people, do not harm animals, do not harm self, and do not harm property) and selects the course of action that results in the least violation of those principles in case of conflict. The system was proven to not perform an unethical action unless the rest of actions available are even less ethical. In contrast, relatively few researchers have been dedicated to bottom-up approaches; an example being the proposal by Wu and Lin of a reinforcement learning agent that integrated human policy, based on the premise that most human behavior is ethical, to accomplish a purpose with less risk of violating the ethical code [42]. There are also some studies that use a hybrid approach to implement ethics by combining top-down

and bottom-up methods, such as Anderson and Anderson proposal of GenEth, a system that analyzes ethical dilemmas through the representation of a variety of aspects of those dilemmas such as the situational features, duties and actions, plus the production of abstract ethical principles using inductive inference before a self-made Ethical Turing Test is used to evaluate those principles by only allowing acts that an ethical expert would accept [5].

For intelligent autonomous computing agents to be fully integrated into society, it is not enough that they have an ethical reasoning, assurances are equally required for these agents to always perform within acceptable legal and social standards. The existing codes of ethics do not reflect the effects of autonomous and intelligent computing agents and are insufficient in terms of legal processes for coping with the inherent risks [4]. Accordingly, the current codes of ethics require a rigorous re-examination to legally regulate these entities. Control mechanisms are essential to ensure that intrinsic laws are always functioning and that specific standards are enforced for the design, development, assessment, use, and maintenance of autonomous agents. Other control mechanisms are needed to inspect and audit the first control mechanisms. Moreover, the identification and assignment of legal responsibility to those accountable for the harm arising from autonomous systems noncompliance with the laws is crucial.

3 Methods

The pillars of contemporary science are here assumed as the conjunction of [Truth, Good, Beauty] that is modeled, respectively, by [Physics, Biology, Economics] of contemporary science. They constitute contemporary science's Absolute, which are hence the best contemporary science has for describing the Transcendent, a pre-requisite for the establishment of ethics, as previously noted by Wittgenstein (1922).

In this work; Physics is a generic term that goes from mathematics as a foundation, through the theoretical physics of gravity and the Standard model, to chemistry as a practical application of quantum physics; Biology is a generic term that refers to the study of "information of life" that goes from the chemistry of organic molecules, through genetic biology, to neuroscience of population groups; Economics is a generic term that goes from the economics of representation in neuronal aggregates, through social psychology and sociology, to the economics of the performance of countries in the midst of the financial policy of a globalized world, all of which are associated with the stipulation of the value associated with transactions; and sociology is described as an aspect of economics, rather than the opposite perspective, because sociology is limited to the description of human social systems whereas economics can easily be extended to other types of interaction [7,19].

To assess the contribution of each of Plato's Transcendentals to the value of an entity/act, meaning the contribution of each of the axes of the 3-axes value, it is important to assess the threefold structure of evolution [28,29]:

- Deterministic Inference → Information from past is preserved into the future thus maintaining survival capability of the entity;
- Natural Selection → Information creation allows new behaviors that alter the resource extraction from environment, the entities with better survival capacity endure;
- Aesthetic Evolution → Alterations of information representation compatibility between entities alter the flow of resources between them, the communication link with better survival capacity endures.

The use of model-free approaches to data analysis is now typically called deep learning, but can also be referred to as machine learning or neural networks, and consists on learning the most effective representation of the data. Deep learning models have been able to show that: most mutations in humans are neither beneficial nor harmful for natural selection until an environment change makes a certain mutation become relevant [31]; specific gene mutations (meaning the deterministic inference of genetic information is partially broken) are associated to metabolism [36]; and Müllerian mimicry does occur in the Darwinian evolution for butterflies [10]. The existence of Müllerian mimicry [10] together with the existence of evolutionarily advantageous characteristics that are not truthful [24] are a further indication that the aesthetic evolution [28,29] occurs and is *de facto* a different evolution line from natural selection.

The aesthetic evolution is a different concept from natural selection, as the strength of the survival is based in the establishment of jointly-accepted symbols [28,29], and not necessarily in a better usage of the environmental resources as occurs in natural selection. The aesthetic symbol might represent a true best-guess of the natural selection trend [10] or not [24], but since natural selection is not deterministic [34] the natural selection trend is indeterminable, and hence its representation by a symbol is always a guess. Moreover, this aesthetic evolution allows culture/ethics/morality to have a certain degree of independence from both deterministic inference and natural selection. The 3-aspects of evolution can be directly linked to the 3-axes of Plato's Transcendentals, Truth+Beauty+Good, by the relation: "deterministic inference"↔Truth + "aesthetic evolution"↔Beauty + "natural selection"↔Good.

Plus, those 3-axes have previously been connected to human neuroanatomy, human decision-making, human mental health, human culture, and human economics [19,20]. The trust-level vs. Gross Domestic Product (GDP)/capita across nations allows for the definition of religion-based clusters [7]. Moreover, across different nations the crime-rate correlates positively with the belief in Hell's existence and negatively with the belief in Heaven's existence [32]. Plus, there is an agreement in [1] and [25] that the appearance of a "Leviathan", meaning a command-and-control hierarchical structure with "stationary bandits" building a political elite and a rule-of-law establishing the rules within the "Leviathan", is a key contributor to the improvement of the wealth of the nations. Moreover, in [14] is described a correlation between the decrease in fear of legal punishment and the reduction of wealth in western nations. Thus, it is an appropriate resume of the relation between wealth and culture to consider that societies are

constituted by a "Tower" that implements the rules of the "Leviathan", and a "Square" where under the protection of the "Leviathan" people can build continually evolving economical interactions [15].

4 Results and Discussion

The relations, respectively, between axiomatic systems and psychology's structure (see Fig. 2) [19], between axiomatic systems and psychology's consciousness types (see Table 2) [19], between the brain 3-axes and brain function (see Fig. 1) [19,20], plus the relation between Plato's Transcendentals [20] and contemporary neuroscience axes (see Table 1) [19,20] allow the clarification of the relation between the 3-components of evolution and the 3 axes.

The relation of Plato's Transcendentals with the brain axes is described in Table 1, but the establishment of the relation with the axiomatic systems described in Table 2 requires the use of the Types of Consciousness of Fig. 2.

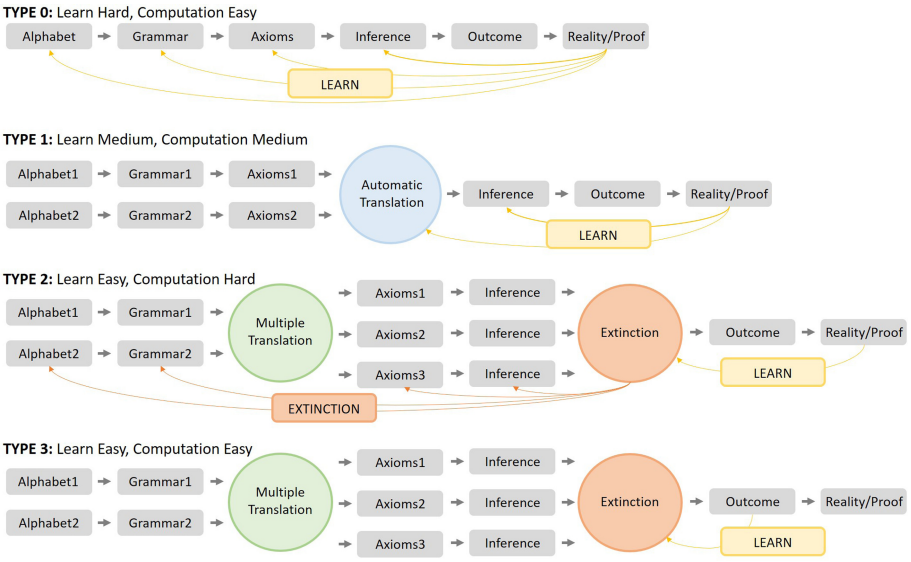

Fig. 2. Relation between axiomatic system, from Alphabet to Output, and the structure of the connection relating the Information-based axiomatic system to the Consciousness Types 0-1-2 described in [33] and Consciousness Type 3 proposed in [19].

The Type 1 consciousness activates the appropriate "deterministic inference" mechanisms so that from a certain stimuli a unique and appropriate response is obtained [33], whereas the Type 2 consciousness activates the appropriate "natural selection" analysis of what would be the consequences of the different

Table 2. Information-based axiomatic system and its relation to psychology [19].

Axiomatic system	Psychology's narrative
$Alphabet(S)$ + $Grammar(P[w], P[w_j], P[< l_{\sim_j} \vert w_j >], \dots)$	Setting
Proof-checking algorithm input (internal vs. external)	Initiating event
Axioms (consistent vs. complete)	Internal response
Rules of inference (single-alternative vs. multi-alternatives)	Goal + Actions
Inferred statements	Outcome
Proof-checking algorithm output (internal vs. external)	Ending

choices and environments so that from a certain situation the best path is chosen [18,33], finally the Type 3 consciousness allows for an analysis of the best options regardless of environmental input so that the best aesthetic option can be achieved in as much freedom as possible [19,33]. Hence, the relation between evolution component, Plato's Transcendentals, consciousness Type, contemporary science, and brain axes is:

- Deterministic Inference ↔ Truth ↔ Type 1 ↔ Physics ↔ Ant.-Post/Hunting/Power/Now-Then;
- Natural Selection ↔ Good ↔ Type 2 ↔ Biology ↔ Inf.-Sup./Eating/Pleasure/In-Out;
- Aesthetic Evolution ↔ Beauty ↔ Type 3 ↔ Economy ↔ Left-Right/Choosing/Meaning/Tone-Word.

The relation just above allows for a direct relation between axiomatic systems, human ethics, and human anatomy; which is a key issue in computer science as machine ethics becomes more and more important, and the implementation of ethical reasoning in intelligent machines is not far off. However, for intelligent autonomous computing agents to be fully integrated into society, it is not enough that they have an ethical reasoning, assurances are equally required for these agents to always perform within acceptable legal and social standards.

It is also obtained a new perspective on the relative importance of the top-down, bottom-up and hybrid approaches to machine ethics. For, the top-down approaches are a "deterministic inference"; whereas the bottom-down approaches are separable between "natural selection" types if the learning is based in the environment, and "aesthetic evolution" types if the learning is based in the interaction with the symbolic representations of the other autonomous agents, the hybrid approach. In practice, just like human ethics is based in all three of the 3-axes value and all three of the brain's 3 axes, the appropriate machine ethics should be based in the three components of evolution, and it will thus be an hybrid approach. Which is reasonable, as the likely source of the brain's 3-axes is the biological evolution's 3-components.

One may ask, and many have done so, what is the best ethics. Nevertheless, the best answer this approach obtains is that there is no answer. The best ethics is not this or that, but rather the permanent search for a better ethics

through comparison, competition, and selection of ethics [7,15,24,25]; provided that those involved are striving to find what is most True and Good [1,14]. Through this balance between Truth and Good, the Beauty is found as a cultural creation of the balance between Truth and Good.

Moreover, for classical ethics, which separate between goodness and evil, to exist in the observed universe, the following is required: i. events universally definable as a goodness constitute a set B; ii. events universally definable as an evil constitute a set M; iii. set S of well-intentioned actions, meaning, they intend to obtain B events; iv. set C of ill-intentioned actions, meaning, they intend to obtain M events; v. classical ethics exists in the universe if and only if: "S actions not intersecting C actions" implies "B events not intersecting M events".

Requirement v means that for classical ethics, the "God/Absolute cannot write right by crooked lines", and if it is valid that "God/Absolute cannot write right by crooked lines" then causality in the observed universe would have to be only by deterministic inference, but that is not the case, as both natural selection and aesthetic evolution also occur. Hence, only a statistical perspective of ethics agrees with the observed universe. Thus, for example, set S actions generate not only set B events with high probability but also set M events with less probability. In rare most extreme cases, set S actions can be so creative that they only generate set B events for all space and time; or set C actions can be so malicious that they only generate set M events for all space and time. Moreover, there may be actions that are selfish, meaning, that they generate B events in their vicinity and M events away from it; or an action can be heroic by generating M events in their vicinity and B events away from it.

5 Conclusion

Alongside AI and Information and Communication Technologies (ICT), AmI has gained a prominent place in the scientific community. As AmI research matures, increasingly superior forms of intelligence and automation are permeating every aspect of human life. Unquestionably, as computers' decision-making roles grow and society increasingly relies on computer-based intelligence with reduced human supervision, ethical considerations are inevitable. In the last years, AmI has experienced a tremendous growth, but few authors have dedicated to the social, moral, and legal implications of this emerging reality.

In this study, we sustain that the success of AmI relies heavily on the development of better ethics and, consequently, on the implementation of effective machine ethics. Hence, the main purpose of the study was to improve the capacity for machine ethics to approach human ethics by assessing the computation of transaction values. We used a 3-axes value approach - Truth+Beauty+Good - to establish a relationship between biological evolution, brain axes, philosophy, psychology, and axiomatic systems.

According to the statistical perspective of ethics, well-intentioned actions are more likely to generate goodness, and malicious actions are more likely to generate evil. Nevertheless, one may ask, why the reason for living should be the

ethics of goodness, and not an alternative ethics of generating evils. The reason for this preference is that what is goodness, is so because it is in accordance with the egalitarian objectivity of the Truth of Physics and with the elitist life-enhancement of the Good of Biology, and hence goodness is what maximizes the occurrence of Beauty; that is, Beauty/goodness is the combination of egalitarianism and elitism that allows for biological life to occur despite physical objectivity. In short, evil in going against life is hence necessarily self-destructive, thus in this aspect Augustine of Hippo [13] is correct in stating that Absolute Evil does not exist, for existing is a goodness, and the Absolute Evil in existing would cease to be Absolute Evil because it had at least one goodness.

Acknowledgments. This work has been supported by FCT – Fundação para a Ciência e a Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

1. Acemoglu, D., Robinson, J.A.: Why Nations Fail: The Origins of Power, Prosperity, and Poverty. Crown Books, Largo (2012)
2. Allen, C., Smit, I., Wallach, W.: Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf. Technol. **7**(3), 149–155 (2005)
3. Anderson, M., Anderson, S.L.: EthEl: toward a principled ethical eldercare robot (2008)
4. Anderson, M., Anderson, S.L.: Machine Ethics. Cambridge University Press, Cambridge (2011)
5. Anderson, M., Anderson, S.L.: GenEth: a general ethical dilemma analyzer. Paladyn J. Behav. Robot. **9**(1), 337–357 (2018)
6. Andrade, F., Neves, J., Novais, P., Machado, J., Abelha, A.: Legal security and credibility in agent based virtual enterprises. In: Camarinha-Matos, L.M., Afsarmanesh, H., Ortiz, A. (eds.) PRO-VE 2005. ITIFIP, vol. 186, pp. 503–512. Springer, Boston, MA (2005). https://doi.org/10.1007/0-387-29360-4_53
7. Beinhocker, E.D.: The Origin of Wealth: Evolution, Complexity, and the Radical Remaking of Economics. Harvard Business Press, Boston (2006)
8. Boughton, J.M.: On the origins of the Fleming-Mundell model. IMF Staff Papers **50**(1), 1–9 (2003). https://doi.org/10.2307/4149945
9. Chaitin, G.J.: Meta maths!: the quest for omega. Vintage (2006)
10. Cuthill, J.F.H., Guttenberg, N., Ledger, S., Crowther, R., Huertas, B.: Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. Sci. Adv. **5**(8), eaaw4967 (2019)
11. Davenport, D.: Moral mechanisms. Philos. Technol. **27**(1), 47–60 (2014). https://doi.org/10.1007/s13347-013-0147-2
12. Dennis, L., Fisher, M., Slavkovik, M., Webster, M.: Formal verification of ethical choices in autonomous systems. Robot. Auton. Syst. **77**, 1–14 (2016)
13. Dyson, R.W., et al.: Augustine: The City of God Against the Pagans. Cambridge University Press, Cambridge (1998)
14. Ferguson, N.: The great degeneration: how institutions decay and economies die. Penguin (2014)
15. Ferguson, N.: The square and the tower: networks and power, from the freemasons to Facebook (2018)

16. Floridi, L.: The Blackwell Guide to the Philosophy of Computing and Information. Wiley, Hoboken (2008)
17. Floridi, L.: The Philosophy of Information. OUP, Oxford (2013)
18. Kahneman, D.: Thinking, Fast and Slow, Farrar, Straus and Giroux (2011)
19. Lori, N., Samit, E., Picciochi, G., Jesus, P.: Free-will perception in human mental health: an axiomatic formalization. from: automata's inner movie: science and philosophy of mind. chapter viii, curado, m., gouveia, ss (2019)
20. Lori, N., Neves, J., Alves, V.: Some considerations on the estimation of the value associated to a clinical act. Procedia Comput. Sci. **170**, 1041–1046 (2020)
21. Lori, N.F., Jesus, P.R.: Matter and selfhood in Kant's physics: a contemporary reappraisal (2010)
22. Machado, J., Abelha, A., Neves, J., Santos, M.: Ambient intelligence in medicine. In: 2006 IEEE Biomedical Circuits and Systems Conference, pp. 94–97. IEEE (2006)
23. Machado, J., Miranda, M., Pontes, G., Abelha, A., Neves, J.: Morality in group decision support systems in medicine. In: Essaaidi, M., Malgeri, M., Badica, C. (eds.) Intelligent Distributed Computing IV, pp. 191–200. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15211-5_20
24. McKay, R.T., Dennett, D.C.: The evolution of misbelief. Behav. Brain Sci. **32**(6), 493–510 (2009)
25. Morris, I.: War! what is it Good For?: Conflict and the Progress of Civilization from Primates to Robots. Farrar, Straus and Giroux (2014)
26. Obstfeld, M., Shambaugh, J.C., Taylor, A.M.: The trilemma in history: tradeoffs among exchange rates, monetary policies, and capital mobility. Rev. Econ. Stat. **87**(3), 423–438 (2005)
27. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books, New York (2018)
28. Prum, R.O.: Aesthetic evolution by mate choice: Darwin's really dangerous idea. Philos. Trans. Roy. Soc. B: Biol. Sci. **367**(1600), 2253–2265 (2012)
29. Prum, R.O.: The evolution of beauty: how Darwin's forgotten theory of mate choice shapes the animal world-and us. Anchor (2017)
30. Rodrik, D.: The Globalization Paradox: Democracy and the Future of the World Economy. WW Norton & Company, New York (2011)
31. Schrider, D.R., Kern, A.D.: Soft sweeps are the dominant mode of adaptation in the human genome. Mol. Biol. Evol. **34**(8), 1863–1877 (2017)
32. Shariff, A.F., Rhemtulla, M.: Divergent effects of beliefs in heaven and hell on national crime rates. PLoS ONE **7**(6), e39048 (2012)
33. Shea, N., Frith, C.D.: Dual-process theories and consciousness: the case for 'type zero'cognition. Neurosci. Conscious. **2016**(1) (2016)
34. Smerlak, M., Youssef, A.: Limiting fitness distributions in evolutionary dynamics. J. Theor. Biol. **416**, 68–80 (2017)
35. Sparrow, R.: Killer robots. J. Appl. Philos. **24**(1), 62–77 (2007)
36. Sugden, L.A., Atkinson, E.G., Fischer, A.P., Rong, S., Henn, B.M., Ramachandran, S.: Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nat. Commun. **9**(1), 1–14 (2018)
37. Voiklis, J., Kim, B., Cusimano, C., Malle, B.F.: Moral judgments of human vs. robot agents. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 775–780. IEEE (2016)
38. Wallach, W., Allen, C., Smit, I.: Machine morality: bottom-up and top-down approaches for modelling human moral faculties. AI Soc. **22**(4), 565–582 (2008)

39. Winfield, A.F., Michael, K., Pitt, J., Evers, V.: Machine ethics: the design and governance of ethical AI and autonomous systems. Proc. IEEE **107**(3), 509–517 (2019)
40. Winfield, A.F.T., Blum, C., Liu, W.: Towards an ethical robot: internal models, consequences and ethical action selection. In: Mistry, M., Leonardis, A., Witkowski, M., Melhuish, C. (eds.) TAROS 2014. LNCS (LNAI), vol. 8717, pp. 85–96. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10401-0_8
41. Wittgenstein, L., dos Santos, L.H.L.: Tractatus logico-philosophicus. Edusp (1994)
42. Wu, Y.H., Lin, S.D.: A low-cost ethics shaping approach for designing reinforcement learning agents. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)