12-12-2022

AI “Ethics by Design”: The Unethical Transfer of an Intransitive Human Characteristic to AI Artifacts

Roxana Ologeanu-Taddei
Toulouse Business School, roxana.ologeanu-taddei@umontpellier.fr

AI "Ethics by Design": The Unethical Transfer of an Intransitive Human Characteristic to AI Artifacts

Roxana Ologeanu-Taddei
TBS Business School
r.ologeanu-taddei@tbs-education.fr

ABSTRACT

This conceptual paper argues that the "ethics by design" approach underlines an unethical idea as inherent human properties (reflexivity, concerns, agency, free will) may be transferred to AI. Instead, I propose a redefinition of ethical actions in AI context while the AI in itself can be described by competency, reliability or safety rather than ethical issues.

Keywords

Artificial intelligence, AI design, ethics, responsible AI, agency, reflexivity.

INTRODUCTION

In the context of their increasing awareness of sustainability and responsibility issues, scholars have turned their attention to the social impact of digital technologies. This concern extends the older interest for design ethics of "man-made artifacts and systems" (Chan, 2018) and "technoethics" (Bunge, 1977).

From this standpoint, AI has been criticized for exerting social or managerial control (Zuboff, 2019; Schroeder & Iatridis, 2021) as well as for providing biased results which underline ethical issues such as privacy, fairness, and discrimination, and further biases (Benbya et al., 2021; Berente et al., 2021; John-Mathews, 2022). Moreover, AI algorithms opens new questions of ethical choices in so-called dilemma situations, such for example the self-driving cars which can be programmed to kill its occupants, rather than to kill the pedestrians in the situation known as the "Trolley problem" (Rakowski, 2016; Chan, 2018). Basically, the question raised by AI design is how to develop AI algorithms which take the right decision (action).

Therefore, some scholars have adopted a pragmatic view, focused on the design and implementation of responsible AI (Trocin et al., 2021; Wang et al., 2021) or its use (Mikalef et al., 2022). The emerging approach of "ethics by design" (Akter et al., 2021; Stahl, 2021) advocates the translation into AI design of the norms related to the impacts of AI (Hayes et al., 2020). In this perspective, authors proposed different principles for the design of responsible AI (Jobin et al., 2019; Peters et al., 2020).

This concern has also grown on the practitioners' side, as several countries have already implemented guidelines for ethical AI. For example, the European Commission published the "Ethics guidelines for trustworthy AI". Also, since 2018, the European Commission has seized the European Group on Ethics in Science and New Technologies: Statement on artificial intelligence, robotics and "autonomous" systems. Some researchers have engaged in developing ethical technologies. For example, Bernd Stahl leads the project called SHERPA (Shaping the Ethical Dimensions of Smart Information Systems), which analyzed the impact of AI and big data analytics on ethics and human rights.

Yet, it is not clear how those principles are effectively translated in practice, as Ryan & Stahl (2020) have already highlighted, and there is no study on the effectiveness of those principles in practice. Moreover, the "ethics by design" approach underlines the assumption of universality and objectivity beyond the context of ethical principles for AI (John-Mathews, 2022) while it does not provide information about how this universality and objectivity can be achieved with respect to the diversity of stakeholders involved in different circumstances and their values as well. A few authors have recently pointed out the role of the context for explainable AI (Lipton, 2018; John-Mathews, 2022) and the difficulty to describe the concept of AI interpretability independently of context as well (Miller, 2019). John-Mathews (2022) defined " interpretability " as "the property of an algorithm that empirically reaches at best some end-user desiderata". Nevertheless, it is not clear how this objectivist and substantialist view of "interpretability" may take into account the variety of users' desiderata in different contexts or for different user groups. John-Mathews (2022) argued that AI designers tend to choose AI explanation with little denunciatory power (meaning presentation of an ethical incident such as gender related biases). Therefore, he proposed to reconcile context-independent ethical principles and context dependent empirical assessment by choosing among two scenarios: the rules-based model or the multiplication of testing scenarios by the AI designers in order to solve ethical incidents (such as biases).

Hence, this recent literature shares a main paradigmatic assumption (Alvesson & Sandberg, 2011) according to the which AI algorithms can have ethical properties and can be good by themselves for users (Miller, 2019). Doing so, it envisions ethics from an independent point of view of the

circumstances (or situation), as objective or universal principles which can be assigned to AI algorithms.

This paper aims to challenge this paradigmatic assumption. I argue that the emphasis on universal principles for the AI design can lead to authoritarian rules which coerce users (Adler & Borys, 1996; Adler & Brodozic, 2022), and therefore become also unethical. Moreover, the "ethics by design" approach allows the transfer of ethics, envisioned as capacity to choose and therefore as an intransitive human characteristic, to AI artifacts. This anthropomorphic transfer is unethical as far as ethics in are a matter of human dignity and emancipation (Donati, 2011; Donati, 2019).

Based Donati's humanistic approach (Donati, 2011; 2020; 2021), I develop an alternative assumption ground about ethics which focuses on the relation between human and non-human artifacts including AI. Hence, I propose to use "ethics" in reference to designers and users of AI and to replace the current "ethics by design" by the usability principles which focus on the design of transparent and trustworthy systems. Therefore, the emphasis should not be made on the proprieties of the technologies in themselves but rather in enabling design and enabling rules (Adler & Borys, 1996) for responsible individuals who create relational goods mediated by digital technologies. Those rules should be based on trust, instead of coercive rules and design which deprive users (employees) from the possibility to understand how to make appropriate decisions (Adler & Borys, 1996) and make them dependent on the IT specialists. Consequently, the main issue how to support users' responsible decisions and right actions as the best action possible in the circumstances (Swanton, 2001) rather than how to develop and implement ethical AI.

ALTERNATIVE ASSUMPTION GROUND. ETHICS AS A MATTER OF REFLEXIVITY

From the virtue-based ethics perspective, a good or right action may be defined as follows: "an act is right iff [if and only if] it is overall virtuous". More precisely, "an act counts as virtuous in respect V (benevolent, generous) iff it hits the target of (realizes the end of) virtue V (benevolence, generosity)"; the "overall virtuous" involves that it is.", which means the best action in the circumstances (Swanton, 2001, p. 45). Swanton (2001) makes a distinction between the right action and the "all right" or good enough action, which is not the best in the situation, and which can blend virtuous and vicious features.

Overall, the right action depends on intrinsic human values (such as benevolence, generosity, justice, honesty etc) which orient our actions. Those values are the concerns, meaning that matter to actors and what their reflexivity works upon (Archer, 2018) and upon contextual factors as well; hence it cannot be defined by exterior principles to the circumstances. Making ethical decision requires then an engagement, in a unique context. For this reason, guidelines to be followed independently of the context and of the actor (subject) making the decision cannot be related to ethics. In addition, an ethical solution involved discussion with all the stakeholders in order to reach a collective final solution. Therefore, the underlining values of ethical decision are not only related to the orientation to others but also to the democratic values which drive a shared solution.

The assumption of ethical proprieties of non-human artifact is criticized by the humanistic accounts of ethics.

Donati's humanistic approach defends values as an irreducible, intrinsic human characteristic which nurtures reflexivity. He criticizes prior sociological theories which share the idea of purposeful actions melted the utilitarian vision of the rational choice theory and instrumental rationality based on self-interest, for being antihumanistic (Donati, 2011). Donati (2020) states that "The human is continually redefined through new expressive, cognitive and symbolic distinctions. Among these, primary value is possessed by aesthetic and moral distinctions, for example, the distinction between what is more or less beautiful, what is good or bad, what is more altruistic or more selfish, and so on." (Donati, 2020).

Donati envisions ethics as symbolic action, "i.e. that orientation towards worth" (Donati, 2011, p.30), which refer to the reflexive criteria justifying the morality of an act. (Donati, 2011, p.30). Reflexivity is thus as at the core of this view, according to which the human emerges in our post-modern societies "as the time/space of a new capacity for choice and aesthetic – expressive reflexivity on the part of individuals freed from the restraints of a strictly structured tradition." (Donati, 2011, p.30).

More specifically, Donati (2011) criticizes the absence of human/non-human distinction, which lead to the attribution of anthropomorphic characteristics to non-human entities and the projection of certain human characteristics onto entities such as technologies and AI "that would deprive humanity of certain of its functions and abilities". For Donati, this human capacity consists in relations "that is the product of reciprocal actions of subjects-in-relation with each other" (Donati, 2011, p. 42). This relationality is an irreducible, intrinsic property of individuals (Donati, 2021, p. 223), and therefore draws the boundary between human and technologies (machines) and lays the basis for a relational humanism as an ethical paradigm about the interaction between humans and technologies. The focus on relationality means that "society is made up of relations in which the distinction between the human and the non-human components can never be obliterated (…)." (Donati, 2011, p. 41).

In Donati's perspective, a humanistic approach means developing relational goods, such as trust and reciprocal cooperation, which differ from instrumental goods in that they are generated by relations and are maintained only anchored in those relations while instrumental goods are transactional exchanges, such as money-based transactions. (Donati, 2019). Relational goods "are reciprocally oriented in a supra-functional sense." (Donati,

2011, p. 42). This means that the absence of reciprocity makes actions only reactive or individual actions which consequently loss their social characteristic. The absence of supra-functional sense transforms actions into operations (or functional actions) performed by automated actors lacking intentionality, losing in this way their human characteristic. Relational goods are inherently ethical, as ethics focus on issues of human rights, social justice, and the balance between altruism and self-interest (Bauman, 1993; Chan, 2018)

A REDEFINITION OF ETHICAL DECISIONS IN AI-HUMAN CONTEXTS

Based on this theoretical ground, I suggest to withdraw the anthropomorphism underlined by "ethics by design" and, instead, to propose alternative assumptions and a redefinition of ethical actions and decisions in AI-human contexts.

First of all, ethical characteristics can be insofar reserved to human actions, considered as the right actions reflexively assessed and oriented towards worth, meaning socially shared values, in order to create relational goods.

Therefore, the assessment and orientation to worth cannot be let only to specific roles such as the "technologist" or the AI ethics officer as the actions based on AI are made by operators or other actors, who need therefore to be engaged in those actions, which means reflexivity and free decision about how to use them in the right way.

This means that coercive rules are unethically so far that they do not allow actors to act responsibly, reflexively, and make the right decision in given circumstances, which require situational awareness. In other words, humans are considered an error issue in the loop (Adler & Borys, 1996) and, ultimately, unethical agents. This vision which enhanced the control of humans limited to the role of operators in executing automated tasks deprives human beings of the core of their dignity, which is the reflexivity and human agency, in making choices (Donati, 2011; Zuboff, 1988; Adler & Borys, 1996).

Furthermore, general principles may lead to autocratic rules, therefore unethical themselves precisely because they are defined from an exterior point of view considering the circumstances under consideration. For example, legitimacy of the stakeholders who design transparency as an ethical criterion to assess AI can be questioned by practitioners on the ground who may consider other criteria or assess transparency in a different way that the prescribed one from the outside.

One may argue that those external, coercive principles are necessary to protect people from harm related to data processes and algorithms working that they do not fully understand, if they are not AI savvy. This objection can be countered by legal rules related to data confidentially and privacy for example, as well as the design of transparent and competent (reliable) AI algorithms which enable (Adler & Borys, 1996) individuals (operators, managers) to make ethical decisions. Those principles are useful but yet cannot be confused with ethics.

Therefore, the AI cannot be considered ethical or unethical in itself but, rather, it can have different properties (such as transparency)

As relational good requires a relation between humans, so far as only human can be free, responsible and engaged in nonhierarchically in value (virtue) oriented actions. This relation can be created between designers, managers in charge of the implementation of AI algorithms and users. Free will in making decisions is at core of the relational goods. Therefore, the goal of an ethical design (oriented toward benevolent and reflexive actions as well as to the creation of relational goods) should be enchasing this free will, that is, managers and users' agency for making decisions. Accordingly, the design and the implementation of AI algorithms should be based on enabling rules (Adler & Borys, 1996) for individuals aging ethically to create relational goods mediated by digital technologies. Those rules should be based on trust, instead of coercive rules and design which deprive users (employees) from the possibility to understand how to make appropriate decisions (Adler & Borys, 1996) and make them dependent on the IT specialists.

Enabling rules for AI design emphasizes on how to support users' responsible and ethical decisions, meaning the best (virtuous) actions possible given the circumstances (Swanton, 2001). Adler & Borys (1996) argue that the internal transparency, the global transparency, the possibilities of repair and the flexibility of the system are the and the flexibility of the usability approach which underlines the enabling design. While internal transparency refers to the transparency of "internal functioning of the equipment or procedure as used by employees" (Adler & Borys, 1996, p. 72), global transparency "refers to the intelligibility for employees of the broader system within which they are working." (Adler & Borys, 1996, p. 72). The repair principle focuses on users' autonomy in repairing the technology: "the ease with which users can repair the process themselves rather than allowing the breakdown to force the work process to a halt". (p. 70). The principle of flexibility emphasizes on the design of the technology "to give advice and make suggestions, and users take the controlling decisions after the system displays the requisite data." (p. 74). On the contrary, in the coercive logic of procedure design, any deviation from standard procedure is considered suspect, which leads to employees' deskilling.

The transparency and explainability of AI algorithms can be labelled in the internal transparency category. Therefore, we note that the current literature on "ethics by design" focuses on this principle of internal transparency and tends to obliterate the importance of the global transparency, of the repair and flexibility categories. Repair and flexibility are directly related to the autonomy given to the users (Alder & Borys, 1996) instead of making them simple operators deprived of free will and overall of

their dignity (Donati, 2011). Users' skilling, which requires training taking into account their work context, are at the core of this usability approach.

Enabling design rules may lay the ground for safe, competent, reliable AI and enabling with whom humans may interact ethically.

THE FALLACY OF THE "TROLLEY PROBLEM" AS AN ETHICAL CHOICE

The "trolley problem" is often cited as an ethical dilemma related to self-driving cars, which can be programmed to kill its occupants pr the pedestrians (Rakowski, 2016; Chan, 2018). Let's examine it from the standpoint of the alternative assumptions of ethical actions related to AI.

Originally formulated by Philippa Foot, the scenario assumes that a trolley runs down a track unable to brake, approaching a fork point. An individual beside the track has the time to reach a lever which can enable her to make the trolley change track. If she does not act, five people (inside the trolley) will be killed; but if she pulls the lever and make the trolley turn, another but single person will die. The problem was exposed to illustrate the conflict between utilitarianism, related to minimizing the total harm (here the number of deaths) and deontology, meaning avoiding doing things that are always wrong (here actively kill a certain person). The "trolley problem" has been promoted to label other scenarios in which a self-driving car has to "choose" a path between two paths, each of them leading to people's death (but they vary according to people's age, or number). (Johansson & Nilsson, 2016). Thus, the question raised would be about the "ethical choices" of the self-driving cars, which would programmed to kill its occupants or the pedestrians (Rakowski, 2016; Chan, 2018).

Nevertheless, several objections can be made against. First, there is no proof that this scenario has ever been face by manual drivers according to any accident reports. The formulation of the problem in itself is artificial in comparison to the driving tutorials and drivers' awareness, assuming that the driver has to keep the control of his/her car (Economic Commission for Europe, Inland Transport Committee, "Convention on road Traffic", done at Vienna on 8 November 1968.) The main goal of a driver is not choosing who to kill, but to avoid harm and, overall, to avoid accidents. Therefore, the issue here is about safety and driver's competency (either for a human or non human) in keeping the control of the car.

Moreover, the requires that the person who acts, or the AI algorithm of the self-driving car, has all the information needed to assess the situation as well as its possible outcomes. It involves also the overall control of the situation (being able to make a choice of the outcomes). This would mean an utilitarian decision (based on the rational assessment of costs and benefits). The point here is again about safety and competency, given that, in real life, accident scenario involves more complexity, according to the circumstances of the accident (for example obstacles, seatbelt use, etc) which makes a situation uncertain. Different scenarios may be possible, with a different probability, and decision making involves taking a risk, still trying to avoid injuries or death. A human or an AI algorithm can be wrong in this assessment, but this is a matter of skills in risk assessment and safe behaviors. Overall, this is not an ethical issue. An ethical choice requires free will, reflexivity and an altruistic orientation towards others. Therefore, an ethical choice would be related to a scenario in which the driver of the self-driving car should its self-interest and ultimately sacrifice himself/herself to save other people. Transferred to the AI controlling the self-driving car, the car should destroy itself in order to avoid harming people. But, even here, this act is not a free choice of the self-driving car, so far that AI does not possess reflexivity, and it does not involve altruism. Instead, it is an operational and tactical action executing a decision made by the AI designer, and it involves safety and security assessment and development.

CONCLUSION

This paper argues that the recent approach of "ethics by design" applied to AI conveys an anthropomorphic vision of AI while ethics involves reflexivity, concerns towards the worth and agency, meaning a free will in tacking action. Therefore, I proposed to distinguish between properties which can be assigned to AI, such as competency, reliability and safety; to AI design, namely enabling rules for design, and to humans as ethical actors when they act in AI-human contexts. The example of the Trolley problem illustrates this distinction. Ultimately, ethical decisions can be only made by humans such as designers but also operators, in a specific context, in which those decisions are motivated by benevolence and, overall, concerns related to others.

REFERENCES

1. Abbasi, A., Chiang, R. H., & Xu, J. JAIS Special Issue on Data Science for Social Good.
2. Adler, P. S., & Borys, B. (1996). Two types of bureaucracy: Enabling and coercive. Administrative science quarterly, 61-89.
3. Ajzen, I. (1988) Attitudes, personality, and behavior, The Dorsey Press, Chicago.
4. Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. International Journal of Information Management, 60, 102387.
5. Archer, M. S. (2018). Bodies, persons and human enhancement: Why these distinctions matter. In Realist responses to post-human society: Ex Machina (pp. 10-32). Routledge.

6. Alvesson, M., & Sandberg, J. (2011). Generating research questions through problematization. Academy of management review, 36(2), 247-271.
7. Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. MIS Quarterly, 45(3), 1433-1450.
8. Bodrožić, Z., & S. Adler, P. (2022). Alternative Futures for the Digital Transformation: A Macro-Level Schumpeterian Perspective. Organization Science, 105-125.
9. Bunge, M. (1979). The five buds of Technophilosophy. Technology in Society, 1(1), 67-74.
10. Chan, J. K. (2018). Design ethics: Reflecting on the ethical dimensions of technology, sustainability, and responsibility in the Anthropocene. Design Studies, 54, 184-200.
11. Cecez-Kecmanovic, D., Galliers, R. D., Henfridsson, O., Newell, S., & Vidgen, R. (2014). The sociomateriality of information systems. MIS quarterly, 38(3), 809-830.
12. Donati, P. (2011). Modernization and relational reflexivity. International Review of Sociology, 21(1), 21-39.
13. Donati, P. (2019). Discovering the relational goods: their nature, genesis and effects. International Review of Sociology, 29(2), 238-259.
14. Donati, P. (2021). Impact of AI/Robotics on Human Relations: Co-evolution Through Hybridisation. In Robotics, AI, and Humanity (pp. 213-227). Springer, Cham.
15. Eurpoean Commission, "Ethics guidelines for trustworthy AI,(" European Commission, Brussels, Dec. 2018. Accessed: Sep. 16, 2019. [Online]. Available: https://ec.europa.eu/digital-singlemarket/en/news/draft-ethics-guidelines-trustworthy-ai
16. Ghani, J. A., Supnick, R. and Rooney, P. (1991) The experience of flow in computer-mediated and in face-to-face groups, Proceedings of the Twelfth International Conference on Information Systems, New York, NY.
17. Johansson, R., & Nilsson, J. (2016, September). Disarming the trolley problem–why self-driving cars do not need to choose whom to kill. In Workshop CARS 2016-Critical Automotive applications: Robustness & Safety.
18. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.
19. John-Mathews, J. M. (2022). Some critical and ethical perspectives on the empirical turn of AI interpretability. Technological Forecasting and Social Change, 174, 121209.
20. Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57.
21. Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side'of AI. European Journal of Information Systems, 1-12.
22. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.
23. Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible AI—two frameworks for ethical design practice. IEEE Transactions on Technology and Society, 1(1), 34-47.
24. Shneiderman, B. (1998) Designing the User Interface - Strategies for Effective Human-Computer Interaction, Addison-Wesley.
25. Stahl, B. C. and Markus, M. L (2021): Let's Claim the Authority to Speak out on the Ethics of Smart Information Systems. MIS Quarterly. Special Issue: Next-Generation Information Systems Theories.
26. Stahl, B. C. (2021). Ethical issues of AI. In Artificial Intelligence for a Better Future (pp. 35-53). Springer, Cham.
27. Suhir, E. (2013). 'Miracle-on-the-Hudson': quantitative aftermath. International journal of human factors modelling and simulation, 4(1), 35-62.
28. Swanton, C. 2001. A Virtue Ethical Account of Right Action, Ethics 112: 32–52
29. Trocin, C., Mikalef, P., Papamitsiou, Z., & Conboy, K. (2021). Responsible AI for digital health: a synthesis and a research agenda. Information Systems Frontiers, 1-19.
30. Tractinsky, N. (1997) Aesthetics and Apparent Usability: Empirically Assessing Cultural and Methodological Issues, *Proceedings of the CHI 97*, Atlanta, GA.
31. Wang, Y., Xiong, M., & Olya, H. (2020, January). Toward an understanding of responsible artificial intelligence practices. In Proceedings of the 53rd hawaii international conference on system sciences (pp. 4962-4971). Hawaii International Conference on System Sciences (HICSS).
32. Zhang, P., Benbasat, I., Carey, J., Davis, F., Galletta, D. and Strong, D. (2002) Human-Computer Interaction Research in the MIS Discipline, *Communications of the AIS*, 9, 20, 334-355