ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.015

ETHICAL DESIGN OF EMBODIED AUTONOMOUS SYSTEMS

M. ANTONIETTA GRASSO and DANILO GALLO
Naver Labs Europe - 6, chemin de Maupertuis
Meylan 38240, France
E-mail: antonietta.grasso@naverlabs.com, danilo.gallo@naverlabs.com
www.naverlabseurope.com

With the arrival of intelligent and learning technologies, embodied or not, new questions have arisen around the best fit of the human activities with them. The study of how to best fit the human with the machine to gain productivity has started as early as the first industrialisation steps. With the advent of networks, computers have gone beyond productivity tools and become support for collaborative activities. With their large-scale learning from data, embodiment and sensing extensions, they are now entities that can be defined as *intelligent*, given their capability of mimicking some and exceeding some other human capability. They are proactive in the sense that they can autonomously take the initiative and they are learning in the sense that their behaviour is continuously tuned by their interaction with the environment. These major changes have induced also changes in the way to study the fitting of the human with the machines. In the beginning, it was a matter of controlled lab experiments where features like human perception, movements and memory were studied to understand how to best shape the interaction to gain efficiency. Subsequently, the object of study has become a complex socio-technical entanglement. Facing the complexity of the new object of study new methods have been adopted capable of grasping a picture much closer to the reality of the new type of interaction. In parallel has come the recognition that multiple actors are affected by the introduction of the machines and tensions are typically present about their interests in using them. This methodological extension has been due to the impossibility of isolating specific variables with the method typic of hard sciences inquiries that had been adopted in the laboratory studies. Lately, the HCI community has undergone another extension: while the first studies were oriented toward productivity gain, new objectives started to appear oriented toward supporting *human flourishing* [1]. The objective of usability of computing is clearly not in contradiction with this new objective, but it is rather an element of a larger picture. This larger picture includes a reflection on what is the impact of putting in place a socio-technical autonomous system. When considering a socio-technical perspective the impact needs to account for a multi-stakeholder perspective and short- and long-term aspects [2]. Indeed, the shift to learning systems introduces new concerns already at the usability level, given that their behaviour could be difficult to understand and the principle of consistency could be easily broken. When moving to embodied intelligent systems a physical component is presents which adds an additional dimension to consider for what could evolve, go wrong and leveraged. Behaviour could be more intrusive, but on the other side physical elements could be leveraged as an additional channel to make clear a behaviour or lack of. A system that is exhibiting a complex and evolving behaviour (i.e. which is *AI-infused*, according to [3]) therefore needs to make use of a number of elements: 1) make clear to the user what are its capabilities, 2) use the sensing to understand the specific situation and adapt to it, 3) recognise when wrong and put in place repair measures, 4) demonstrate how it is evolving [3]. These guidelines have been defined upon examination of more than 20 years of studies and the direct examination of 13 products [3]. This is the minimal set of requirements that usability demands for this new generation of technology, embodied or not. However, as mentioned earlier, the more powerful and adaptive behaviour this technology exhibits, the more an examination of its impact is demanded, to avoid harm and possibly produce individual and societal good. Technology has never been as much under scrutiny as recently, with a number of guidelines and assessment methods being

disseminated. Ethics is about humankind identifying what values are fundamental and worth defending. This can be done in a protective way to avoid harm, or in a propositive way to promote the good. Additionally, values that are contextual, i.e. more specific to a certain group of stakeholders, are also present. An example of the first type is the EU guidelines HLEG [4], an example of the second type are the UN guidelines for sustainable growth [5], while contextual values can be derived case by case with methods like Value Sensitive Design [2] and an example is given in [8].

While guidelines about the proposed values have been defined, the work about mapping those to technical solutions remains still scattered and highly fragmented in various subdomains. We have done a preliminary review of the body of work and identified three subcategories of area of intervention. The first pertains to the collection and use of the data. The new statistically based models require large datasets while sensing, and user implicit and explicit feedback can fine-tune and enrich the system functionality. However, how these data are collected and used requires user awareness. Additionally, in the case of foundational models, these data can originate new data that can be biased and present toxic and harmful elements. The second major area of intervention relates to the autonomy that is left to the human in face of the intelligent capabilities of the systems. Various aspects of ethical concern that need to be addressed exist. In the case of manipulative systems using *dark patterns* [7], the question is how to leave control to the user when e.g. connecting to social media. At work, there could be issues of deskilling caused by the replacement of human capabilities e.g. when utilising diagnosis systems and also issues of human dignity when technology is enacting algorithmic management over the employees and deciding the order, length and tools they have to use to perform their tasks. Finally, a third category relates to how to support the exhibition of behaviour that can be judged as trustworthy. Trust can be defined as exhibiting consistently over time a competent and not harmful behaviour. While usability issues certainly relate to this topic, it pertains to ethics the design of technology that is not causing harm and prevents confidence loss even if aimed at positive values. Also, it is worth noticing that these areas of intervention are correlated and any analysis of them in a specific system should consider them holistically to identify possible tensions. While in the case of services, design guidelines have been put forward, for autonomous embodied systems the effort is still spread across specific projects and application domains. We believe that a systematic effort is required to start to frame existing design solutions and propose associated patterns and design methodologies. This is an effort we are currently undertaking.

References

1. Paul Dourish. 2019. User experience as legitimacy trap. *Interactions* 26, 6 (November - December 2019), 46–49.
2. Batya Friedman and Peter H. Kahn. 2002. Human values, ethics, and design. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. L. Erlbaum Associates Inc., USA, 1177–1201.
3. Saleema Amershi et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19*). Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13.
4. HLEGAI. (8 April 2019). High-Level Expert Group on Artificial Intelligence, EU - *Ethics guidelines for trustworthy AI*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
5. United Nations. Transforming Our World: *The 2030 Agenda for Sustainable Development*. Pub. L. No. A/RES/70/1 (2018). United Nations.
6. Luciano Floridi et al. How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics* 26, 1771–1796 (2020).
7. Arunesh Mathur et al. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Association for Computing Machinery, New York, NY, USA, Paper 360, 1–18.
8. Tommaso Colombino et al. 2021. Ethical Design of a Robot Platform for Disabled Employees: Some Practical Methodological Considerations. *Frontiers in Robotics and AI* 8:643160.