Non-human agencies: a twofold perspective

Guglielmo Papagni, MsC, guglielmo.papagni@tuwien.ac.at
Institute for Management Science, TU Wien

Introduction

Artificial agents, in the broadest interpretation of the term, are becoming more and more present in our lives. In some cases their intervention is evident (for example with self-driving cars or social robots), while in others they can influence our lives almost without us recognizing it (e.g. content recommenders, hiring algorithms).
Several questions progressively arise, as the number of artificial agents increase, as well as their range of application. Since we can expect this trend to grow in the coming decades, and the different dimensions to become even more mingled (e.g., embodied forms of AI), the need to address these questions grows accordingly.

Within the **EPSRC Principles of Robotics**, the issue of robots' transparency is directly addressed: “Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent” [1]. We can extend this assumption also to other types of artificial agents, and approach some of the critical issues that are implied in the idea of transparency.

Stemming from the previous description, this research project aims to approach artificial agents from a twofold, but nevertheless intertwined, perspective. Following the path highlighted by the quote from EPSRC, part of the focus is thus on the “machine nature” of the agents, while the other main issue is represented by the idea of “transparency”.

The fixed center around which the two subtopics orbit is the **human perception** of the interaction, following the idea that it doesn't really matter how we design artificial agents, if there is no acceptance of them on the side of the users.

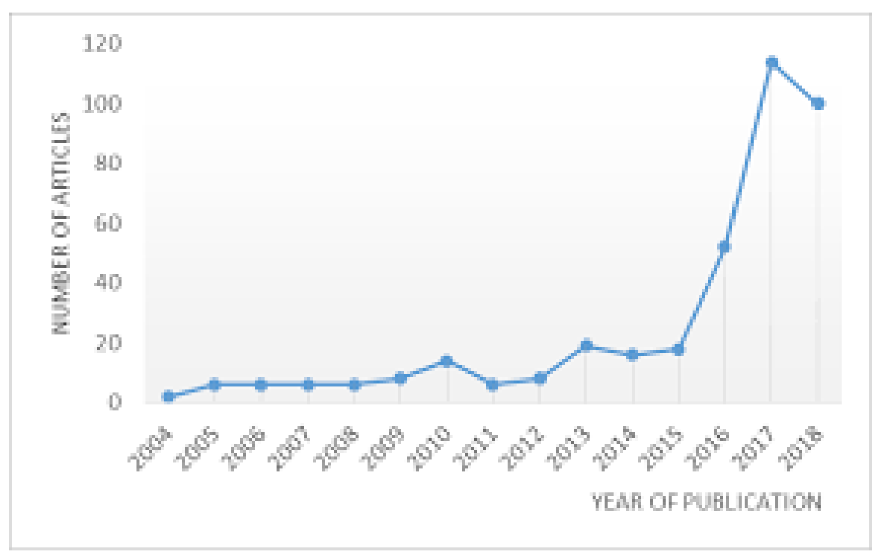
Research questions

This dissertation will mainly address two research questions, both focused on how humans perceive artificial agents.

The first one refers to the recent growth of the research fields of explainable artificial intelligence (XAI), and interpretable machine learning [2,3], and is related to the increasing opaqueness of AI technologies (the figures below show features, i.e., keywords and number of publications of this growing trend in the last years. Source: [3]).

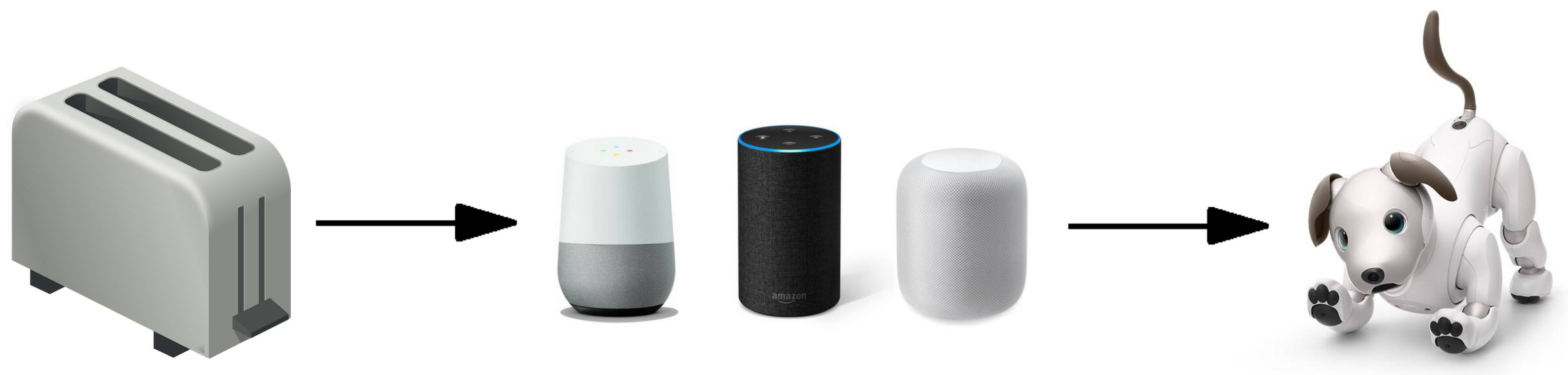
- **How can we be sure that an artificial agent displays the right form and degree and form of transparency and accessibility to gain the user's trust and avoid deception?**

Understandable AI
Comprehensible AI
Accurate AI/ML
Transparent AI
Black box
AGI
Third-wave AI
Interpretable ML
Cognitive science
ASI
Data science
Intelligible ML
Responsable AI
Interactive AI
Explainable AI
Ethics



The second research question arises from the fact that, despite all types of artificial agents should be transparent about their (ontological) nature, in certain cases people perceive them as something in-between an artificial and a living being [4,5] (the images below show how, conceptually, the ontological transition from “fully artificial” to “in-between” is supposed to occur among different kinds of objects).

- **How do people perceive different artificial agents in ontological terms? What kind of impact can this have on a societal level and on interpersonal relationships?**

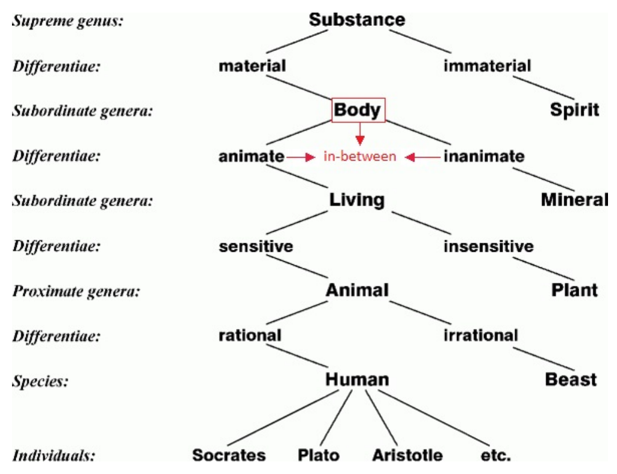


Methodology and work in progress

Both topics are investigated in two phases, one more conceptual and theoretical, and the second empirical. The reason for this approach is that the debates over the concepts involved (i.e., explanatory interaction and ontological categories) are based on long traditions of thought in social sciences, which abundantly predate the arise of artificial agents (e.g., the topic of causal attribution, which is fundamental in understanding explanatory interactions, has a tradition that dates back to Aristotle [2].

In the first case, the theoretical paper (submitted) is aimed elaborating metrics, adapted from human interaction, for improving the design of explainable artificial agents and for questioning concepts related to transparency and **accessibility of the decision making process (DMP).**
The empirical phase contemplates testing in HRI/HCI scenarios some of these metrics, with specific attention for some that so far didn't have enough relevance in the field, e.g., multi-modal communication, iteration of the explanatory act [6,7].

Concerning the second topic, the theoretical paper (work in progress) aims to analyze the question of the ontological status of artificial agent within a strongly interdisciplinary framework (i.e., how HRI/HCI results can contribute to a broader philosophical and social debate).
In empirical terms, the theoretical results are translated into the generation of HRI/HCI scenarios where the boundaries of ontological categories are tested for different types of artificial agents, and the reactions evaluated for a further theoretical generalization.
(The figure on the right shows an “adapted” version of a classical ontological tree, highlighting the point where the variation would be introduced).



Conclusion

The main goal of this project is to shred light on the nature of the interaction between humans and artificial agents, following one of the most basic definitions of ontology, which is the discourse about the nature of beings and of the relations that they entertain with each other. For how some aspects of this investigation might sound speculative (e.g., we still don't have fully autonomous, truly intelligent, embodied artificial agents), other ones are not anymore only a possibility, but are already happening, and we should address them now to try to be ahead of the issues, rather than chasing them [3].

These technologies promise to have a progressively deeper and broader impact on every level of our lives. It is not always clear whether this will have positive rather than negative effects. It is not clear because the role that this technologies will play is not completely defined yet. Thus, being aware of this potential, we should invest our research efforts in understanding how to design them in order to maximize the societal benefit.

Thus, to achieve such an ambitious goal we should in the first place understand how artificial agents are already influencing our society today, and how people react to their introduction, and adapt their design accordingly.

References

[1] Theodorou, A., Wortham, R. H., Bryson, J. J. (2016). Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots. Paper presented at AISB Workshop on Principles of Robotics, Sheffield, UK United Kingdom.
[2] Miller, T. (2018). Explanation in Artificial Intelligence: Insights from the Social Sciences. In Artificial Intelligence, Volume 267 (pp. 1-38). Doi: 10.1016/j.artint.2018.07.007.
[3] Adadi, A., Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access. (pp. 52138-52160). Doi: 10.1109/ACCESS.2018.2870052.
[4] Severson. R. L., Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. In Neural Networks, Volume 23, Issues 8-9 (pp. 1099-1103). Doi: 10.1016/j.neunet.2010.08.014.
[5] Alač, M. (2015). Social robots: Things or Agents? In AI & SOCIETY, Volume 31, Issue 4 (pp. 519–535). Springer London. Doi: 10.1007/s00146-015-0631-6.
[6] Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K. (2019). Explainable Agents and Robots: Results from a Systematic Literature Review. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19).
[7] Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Paper 582, 18 pages. Doi: 10.1145/3173574.3174156.