Workshop Proposal
Ethical Issues of Open Ended-Learning in Autonomous Robots

Daniele Caligiore, Vieri G. Santucci, Gianluca Baldassarre
Institute of Cognitive Sciences and Technologies
National Research Council of Italy
Rome, Italy
Email: {daniele.caligiore, vieri.santucci, gianluca.baldassarre}@istc.cnr.it

I. Concept

By definition, open-ended learning robots might develop behaviours, motivations, goals, and intentions that cannot be anticipated at design time. Beside being a paramount enterprise for the development of truly autonomous artificial agents, open-ended learning poses important practical and ethical issue regarding the future of robotics [1]. Since the presence of artificial agents in our daily life will grow in the near future, how can we manage our interactions with autonomously developing machines that might be driven by both extrinsic and intrinsic motivations, might change indefinitely, and might even acquire an intelligence similar to ours?

A first answer to this question could constrain the developmental process of the robots, preventing them from manipulating possibly-dangerous objects, avoiding some kinds of interactions and, in general, reducing the autonomy and the curiosity of the artificial agents. This strategy could provide some benefits to the application of artificial agents in some real-life situations, in particular those where we properly know all the elements that play a crucial role. However, the real world, and in particular the human environment, has a complexity that makes it virtually impossible to prevent all the potentially critical situations that the robots could autonomously discover. Indeed, facing challenges unexpected at design time is a major motivation behind developmental robotics and constraining the learning and the development of artificial agents would seem a dramatic distortion of the fundamental ideas behind open-ended learning. So, how to find a trade-off between the development of autonomous robots and the necessity to regulate their behaviours?

A possible pathway, suggested by the developmental robotics approach, could be to implement solutions as those employed in human development, where children's learning is open but at the same time strongly constrained by the specific environment where they are raised: from the initial guidance by the family to the influence of the social environment that gradually regulates all aspects of their life - the friends circle, the school, the ethical values, the social norms, and laws.

Following this pathway, an important open question is the possibility of distinguishing between different levels of cognition. At first sight, sensorimotor autonomous exploration seems safer if the robot operates in a safe environment (e.g., it can handle objects not too fragile). As for the babies, in this case the acquisition of a proper sensorimotor control can be considered a process where the greatest risk is for the robot itself. On the other hand, the application of the acquired skills to pursue more complex goals seems to require higher-level evaluations that may imply social/ethical/moral knowledge. However, even this distinction presents some flaws: what if a robot, to test the physical properties of objects, tries to explore "the delicate vase of grandmother", or to "manipulate" a human being, or to play with objects dangerous for it or for humans such as a sharp tool or a vehicle?

Moving to a different level, the increasing presence of autonomous robots in many aspects of everyday life (work, transportation, free-time, etc.) have raised legal issues related to artificial agents. For example, the European Union started discussing the possibility to assign a legal status to artificial agents [2]. This is particularly relevant for the case of open-ended learning: who is responsible for behaviours that the robot might acquire after it has come out of the factory? And what if the robot is "trained/biased" by a human user in ways that end up damaging other humans?

For the future, even more complex problems are expected. The first is the "psychological threat" represented by the fact that open-ended learning robots are expected to acquire an increasingly complex intelligence that might possibly achieve and even overcome human intelligence. Indeed, robots undergoing a developmental process that is increasingly similar in quality and complexity to the one of humans (the objective of developmental robotics) might arrive to acquire emotional and cognitive features similar to ours. This poses a psychological threat for humans, that makes us wonder what it means to be human once the boundaries separating humans from robots, and related to complex emotions, higher level cognition, self, and even consciousness, get increasingly thinner [3]. Beyond this, the workshop discussion might also look far into future possible "existential risks", namely scenarios where robots that get increasingly intelligent and autonomous might arrive to decide they do not need humans anymore [4].

II. Target audience

To discuss these important topics we propose a full-day workshop with speakers coming from different disciplines, with keynote presentations and a final round table. The target audience of the workshop will be researchers interested in developmental robotics and open-ended learning, and at the same time concerned on the implications that the technologies they produce might have for the future well-being of humanity.

III. List of speakers

- Minoru Asada (Osaka University, Japan)
- Gianluca Baldassarre (National Research Council, Italy)
- Daniele Caligiore (National Research Council, Italy)
- Benjamin Kuipers (University of Michigan, USA)
- J. Kevin O'Regan (Universit Paris Descartes, France)
- Pierre-Yves Oudeyer (Inria, France)
- Daniel Polani (University of Hertfordshire, UK)
- Matthias Rolf (Oxford Brookes University, UK)
- Elmar Rueckert (Technische Universitt Darmstadt, Germany)
- Vieri Giuliano Santucci & Daniele Caligiore (National Research Council, Italy)
- Guglielmo Tamburrini (University of Naples "Federico II", Italy)
- Toby Walsh (University of New South Wales, Australia)

Acknowledgment

Supported by the EU FET Open project GOAL-Robots - Goal-based Open-ended Autonomous Learning Robots n. 713010

References

[1] K. Schwab, *The Fourth Industrial Revolution*. UK: Penguin, 2016.

[2] R. Viola, *The future of robotics and artificial intelligence in Europe*. European Commission (Blog Roberto Viola), 2017. [Online]. Available: https://ec.europa.eu/digital-single-market/en/blog/future-robotics-and-artificial-intelligence-europe

[3] Future of Life Institute, *Benefits and risks of artificial intelligence*, 2016. [Online]. Available: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/

[4] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press, 2014.