Ontological challenges to Cohabitation with Self-taught Robots

Stefano Borgo
Laboratory for Applied Ontology, ISTC, Consiglio Nazionale delle Ricerche, Italy
E-mail: stefano.borgo@cnr.it

Editors: Pascal Hitzler, Kansas State University, Manhattan, KS, USA; Krzysztof Janowicz, University of California, Santa Barbara, USA
Solicited reviews: Christopher Schlieder, University of Bamberg, Germany; Boyan Brodaric, Geological Survey of Canada, Natural Resources Canada, Ottawa, ON, Canada

Abstract. When you meet a delivery robot in a narrow street it stops to let you pass. It was built to give you precedence. What happens if you run into a robot that was not trained by or for humans? The existence in our environment of robots which do not abide by human behavioral rules and social systems might sound odd, but is a case we may encounter in the near future. In this paper, self-taught robots are artificial embodied agents that, thanks for instance to AI learning techniques, manage to survive in the environment without embracing behavioral or judgment rules given and used by humans. The paper argues that our ontological systems are not suitable to cope with artificial agents. The arguments are speculative rather than empirical, and the goal is to drive attention to new ontological challenges.

Keywords: Ontology, Robotics, Interaction, Multi-agent systems, Heterogeneous agents

1. Introduction

The marriage between robotics, even when enriched with Artificial Intelligence (AI) techniques, and semantic approaches has not worked well in the past. For long time roboticists have been concerned with hardware limitations and search problems in core areas like kinematics control, navigation, object recognition, obstacle avoidance and so on [1, 2] At that time, the information stored in the robot used to be carefully selected, and encoded in *ad hoc* fashion, to optimize reasoning for those tasks. The use of knowledge as a (conceptual) tool along with actuators, like a gripper, and sensors, like a camera, was a luxury limited to envisioned future scenarios.

Now things have changed. In the last 10 years the continuous improvement of hardware and information processing capabilities has led roboticists to imagine general purpose agents acting in open environments. This vision requires to develop planning techniques for multiple coexisting goals that go beyond traditional robot's control and navigation, and to build robots that can reason in terms of actions, plans, expectations, other agents' intentions and possible collaborations. Important technical and conceptual consequences were brought up by this change like, e.g., the distinction between geometric planning and task planning [3], and the distinction between behavior (roughly, how the agent interacts with the environment) and function (that is, how that behavior contributes to the achievement of a goal) [4].

Once the need to enrich the robot with models for environment, goals, actions, functions and behaviors became clear, the community started to investigate suitable semantic approaches, e.g., [5, 6]. Today semantic techniques and applied ontology methodologies are largely exploited for a variety of tasks like decision making, belief update, situation assessment, interaction and communication. Interest in ontological modeling is further witnessed by the release of a dedicated standard in the area of robotics and automation [7].

When the aim is to develop general purpose autonomous robots, the information system of the robot must be able to process, integrate, store, recall and update information coming from a variety of sources (e.g., different kinds of sensors as well as different types of collaborators) and deliver dedicated information based on goals, detected environment and decision making processes. Furthermore, the knowledge model, which uses this information to build a view of the environment, must also elaborate possible outcomes, detect the presence of other agents and relevant objects, predict other agents' goals and future actions. Ideally, such a system takes advantage of techniques in information science and uses ontology as pivot to ensure the reliability of the information management. This view looks very promising today since important limitations, like memory capacity and processing speed, have been largely removed.

What is the role of applied ontology in this setting? Applied ontology has been introduced to overcome interoperability problems, primarily at the semantic level, caused by the existence of different perspectives, e.g., databases developed by different organizations or interpretation mismatches by agents with different roles. What pushed the ontologists to believe in the possibility of information integration was the simple observation that all information is about reality or about human views of reality, since information (via human perception or human designed sensors) as well as its interpretation is human-based and since "reality cannot be (self-) contradictory", as the motto goes, as long as each agent is locally consistent, everything can be managed to fit. Of course, we have general ontologies that make incompatible choices and are mutually inconsistent. Yet, it is assumed that humans can, at least in principle, understand each other's ontological system, and even switch from one system to another as needed. Practically, two human agents relying on different ontologies may need to go through an interaction phase to understand each other's viewpoint but at the end they can correctly interpret the information they exchange. Or so it is believed. How to formally model this from the logical viewpoint remains a problem due, so the assumption goes, to the limitations of today's formal ontology understanding and of the adopted logical systems.

In other domains do not need to make these assumptions. For instance, in the semantic web view ontologies model circumscribed interests, and do not make fundamental claims about reality. If some form of information integration is needed, alignment, extension and mapping are the actual targets, anything more being a plus. Unfortunately, this latter approach is not sufficient if the goal is to integrate the views of humans and robots for day-to-day cohabitation, possibly enhancing collaborations and social relationships. To be reliable, the ontology has to model how these agents understand reality.[1] Standard arguments in defence of the existence of a unifying ontology or of the mutual understandability across ontologies rely on the assumption that understanding is human-based. In a world where human and artificial agents coexist and are independent, understanding ceases to be human-based and human-centred. In such a world, can we still defend the existence of a unifying ontological system? Can we still believe in the possibility of mutual understanding at the ontological level? I doubt it.

Clearly, the environment is the same, and so is the material world. Or not? An embodied agent learns how the material world is by establishing relationships with the outside world: it learns and explores the environment via its body, cognitive capabilities, sensors. Robots lack cognition as humans know it, and their body is not only equipped with different sensors, it is not even biological. This makes robots' experience of space, time and matter much different from that of humans. If the understanding of reality (whatever that means) depends on sensing and information processing, as we argue, different types of agent very likely develop different understandings of reality. How much different? That depends on the type of robot we are talking about and is, generally speaking, a complex question since it is unclear how to set up a possible comparison. Indeed, the initial claim that humans and robots are part of the same and only reality now sounds less reassuring.

What we are suggesting is that distinct agents (biological, artificial, cyborg) naturally develop distinct ontologies about what reality is, and that the belief that different agents may agree on a unifying view, or even understand each other's system, is not supported by solid arguments. We should take seriously the possibility that humans and robots act according to views of reality that are not only incompatible but also largely incommunicable. The paper posits this as a problem for the future of our species and societies, and indicates directions where ontological investigation is needed.

[1] In the paper the term 'understanding' has a broad sense as it depends on the agent type. It covers notions like 'building a model', 'attributing a behavior' as well as 'giving meaning' to something.

The rest of the paper is organized as follows: section 2 introduces the distinction between situation and scenario, and characterizes the expression 'self-taught robot'; section 3 focuses on the change of perspective brought into applied ontology by the coexistence of highly heterogeneous robots; section 4 shows the need to discuss interaction in a broader setting; and the concluding section, section 5, points to a variety of related problems addressing briefly the world of non-human animals.

2. Situations, Scenarios and Self-taught Robots

Due to the ongoing development of robotics, humans need to learn how to cope with robots. This is having an impact on human social behavior, and in particular on conventions [8], that at the moment seem discussed in terms of heteromation [9]. The norms that determine the organization of the human community and of human everyday interactions will adapt to emerging forms of robotics. However, AI methodologies like deep learning and hybrid knowledge-neuronal systems, make possible the creation of robots which achieve autonomy independently of human intervention, and this pushes one to imagine a variety of hypothetical situations. In this paper we are interested in situations that arise in a state of coexistence between humans and robots, the latter understood as embodied artificial agents. In particular, we look at situations in which humans and robots maintain their substantial independence and face the need to share space and resources to achieve their goals (survival, satisfaction of desires, ensuring safety conditions, adapting the environment to special needs).

Let us call *(material) situation* the layout of the world in which the agent(s) is. The situation is essentially the spatio-temporal fragment of the world in which an agent can use its sensors to perceive and its actuators to act. We make a distinction between the situation, which we take to be a state of affairs (actual or possible), and the interpretation of the situation, which depends on the agent, and call the latter *scenario*. For instance, a situation could be an enclosed area with several acting objects sitting in front of another, the latter emiting loud sounds in their direction for some time. This situation can be interpreted in different ways, each identifying a different scenario. One scenario is that an academic seminar is taking place in the room, another that the group is doing a rehearsals of a theatre piece, a third that these are zombies controlled by aliens, etc. Given a situation, humans use several types of information to choose a coherent and socially acceptable interpretation [10]. With the full development of robots we can imagine that this interpretation agreement should include robots, at least those involved in collaborations, and the achievement of a shared interpretation among humans and robots in general is clearly more complex.

Let us call *self-taught robots*, self-robots for short, robots that learns to cope with the environment without human help, that is, without humans having the capability or even the opportunity to control how they learn or how they evolve.[2] In the rest of the paper we focus on self-robots since they are particularly challenging to ontological modeling. Note, however, that self-taught robots, or even self-engineered robots, are extreme cases that we use for exemplification. Our observations are more general and apply also to robots that, while built and trained by humans, are not completely transparent to them, like today's robots based on neural network architectures. The behavior and goals of self-robots is likely neutral with respect to human behavior and interests, and thus far from our usual way to interpret situations. On top of this, for humans self-robots' behavior can be largely unexpected since humans and robots lack even common biological needs (an important source of clues in animal studies). It is reasonable to think that a self-robot, since driven by its sensors, actuators, reasoning capabilities, body organization, and learning experiences, may focus on different elements in attempting to interpret a situation. Any interaction from the side of the self-robot will be based on its own scenarios, and the rules and expectations it has developed with them.

Let us further constrain our focus by concentrating on self-robots that develop similar understanding of the environment, similar behavior and comparable goals provided their hardware is similar and, with respect to this, they live in comparable environments. Let us assume also that these self-robots are fairly coherent in matching behavior, scenarios, goals and actions. In different words, let us restrict our attention on self-robots that manifest regularities in dealing with the environment. This does not go as far as implying that, given a situation and the type of self-robot, they are predictable. The assumptions aim to ensure

[2] This should not be surprising since artificial agents have already the capability to learn how to cooperate via coordination in simple unsupervised settings [11].

that it should be possible to develop some kind of shared set of rules and expectations among humans and self-robots, a set that can be the core for a system of norms [8] with which to regulate cohabiting and, perhaps, sustenance and cooperation among humans and self-robots.

3. Ontologies vs. Agent's Ontologies

Compared to humans, which have very similar bodies and capabilities, robots can be very heterogeneous: they can comprise different sensors and actuators, some even tailored to specific information or actions, can have central or distributed processing units, can process information locally, at the central level or in dedicated subsystems, can reason with different computability and memory resources, and apply deduction rules, default rules, optimisation evaluations, probabilistic assignments, learned associations and so on, perhaps mixing several of these. As we saw, this variety suggests that self-robots can understand a situation in unique ways, but since they show regularity in their behavior, it is reasonable to ask what kind of ontology may such robots develop.

This question is not new. It rephrases the classical interoperability problem for which applied ontology has been introduced, that is, (*a*) to make interoperability possible at the semantic level, and (*b*) to semantically integrate information of different kinds and from different sources. But is ontology engineering, as known today, suitable to model ontological systems not developed by humans?

To cohabit with self-robots and perhaps interact with them for information exchange, collaboration, or simply to avoid ending up in dangerous situations, one needs to build a model of how robots understand reality and what needs they have, and possibly vice versa. The issue is not about merging or aligning ontologies in abstraction, for which there are different techniques in ontology engineering. The issue is whether today's state-of-the-art in ontology engineering can make sense of ontological systems that self-robots could develop (leaving aside the related problem of how to elicit such ontologies).

To understand a self-robot which is physically and conceptually different from humans, that collects and classifies things in the environment according to viewpoints humans do not use or even consider, do humans need broader top-level ontologies (TLO) and a larger spectrum of agent-level ontologies (ALO)? While a TLO can be understood as a foundational ontology in the usual sense, a ALO is here seen as an agent's specialization of a TLO. Thus, if the TLO is primarily influenced by the robot's capabilities (e.g., to sense, reason and act), the ALO is the consequence on the ontological needs that the agent develops because of its history of interactions with the outside world.

What would be an ontology that is not already covered by today's TLO? A simple one, compatible with a robot that senses the environment at regular intervals, would claim that no event exists. There are only scenarios in which knowledge of objects and their properties, like relative position, is regularly override by the next sensing activity of the robot. This robot can develop a notion of causality, e.g., out of regularities in experienced sequences of scenarios. A notion of continuous change might be unaccessible to it (this depends on the relationship between sensing frequency and stimuli processing speed, compare the case of movie frames and vision processing in humans).

In another case the robot's ontology may concentrate on material properties (density, thermal energy storage, chemical stability, etc.) disregarding features usually relevant in human TLOs like shape or even relations like connection. For instance, a robot that has sensors to detect compounds of fluorides, chlorides, nitrates, hydrates and so on, could develop an ontology based on these distinctions becoming a sophisticated version of the biological substances that behave depending on the chemical concentrations in their environment.

These two cases are not really challenging. The first ontology, which to the best of my knowledge has not been formalized as a TLO in applied ontology, presents a reasonable view which is not seriously considered in our culture but is nonetheless at the core of some engineering devices. The second TLO is more likely a fragment of several existing TLOs. For instance, the DOLCE ontology [12] could be extended to include such an approach as an extension module of the Amount of Matter class.

More challenging for todays TLOs are the context-dependent ontologies. Here is an example. Assume a self-robot uses its sensors to check if the environment is as desired. Whenever it discovers a difference, deemed relevant, between the sensor data and the desired state, it solves the problem by building whatever is missing and dismantling whatever is not matching the goal state. The point is that the behavior of this robot is centred on a contextual classification of what it detects. The agent's ALO has at least three classes: a class of goals, a class of things to destroy, and a class of things to create. These classes are not ontological in our usual sense as they all depend on the state of the robot: situation interpretation and goals can change over time while what to destroy and what to create may depend on how the robot checks the environment and even on its position. This ALO is orthogonal to human TLO since no ontological class is admitted, everything is based on interpretation. For instance, the class of things to create could collect entities that in human TLO are in disjoint classes like physical objects, individual qualities, relationships across objects and even arbitrary combinations of these. Note that the robot may not be able to make sense of our basic ontological distinctions since its actions are very strict: to build from scratch and to dismantle. Note also that the data collected in a situation could correspond to the goal or not depending on the quality of the sensors and on the robot's position at that moment. In short, this robot may behave rationally but its view of reality would not be compatible with a TLO in today's terms.

These examples of self-robot ontologies are quite simple and only aim to show unusual cases. The extent of human TLOs is challenged when investigating robots based on neural network approaches since they may understand reality by detecting regularities in flows of information that overcome human capacities.

It might turn out that applied ontology, as we know it today, is fitted to develop TLOs suitable for self-robots, perhaps leaving out only some extreme cases, but most likely this is not so. We need to investigate methodologies that enable understanding of different ALOs, to integrate them within a larger class of TLOs, and to develop interfaces for information exchanges across these TLOs. The results of this line of research help to model the interaction between heterogenous agents like humans and robots, the focus of this paper, but also between different robot types and, to stretch it even further, between humans and aliens.

4. Interaction

This paper started with a practical motivation, expressed via hypothetical situations and scenarios, aiming to drive attention to the new ontological issues we are likely to face. The motivation, to foster possible interactions among human and artificial agents, points to another ontological problem: how should we understand the ontological notion of interaction?

The social systems that humans have experienced so far are either systems that evolved with humans, like cities, or are largely controlled by humans, like farms and hunting scenarios (with or without the support of non-human animals). These social systems are thus human-centred, and humans have been by far the most powerful agent in them. In the hypothetical world we are considering, humans may not have this special position. And this has important consequences.

The question that arises is not whether interaction in these hypothetical cases is possible, but in which sense it is. After all, one may doubt that humans and self-robots can make sense of each other's behavior and expectations. Here the very use of the term 'interaction' may be challenged as it implicitly suggests some kind of purposefulness (at least from one of the interacting entities) combined with some form of reciprocity. Indeed, this is the standard understanding in robotics. In this sense, it brings to mind intentional agents. Even though agents capable of making decision and having goals may fit this view, it seems to me that such purposefulness and reciprocity is too restrictive.

I like to start from the notion of interaction as used in physics and engineering, essentially the way entities influence or affect each other's behavior. Roughly, an interaction described at the level of physical laws, say the interaction between a book and the table where it lays, states that the behavior of one entity is influenced by the presence of the other. This physical interaction can be the starting point for an ontological analysis. Of course, not all interactions apply to entities controlled by physical laws only. A cognitive agent detecting an object in its environment, e.g., perceiving a car on its path, moves its attention to it (manifesting an interaction at the cognitive level) and may change position (interaction at the planning, functional and acting levels) to avoid contact. Other types of interaction occur at the social and cultural levels when joining a queue at an office or singing 'Happy birthday' at a party. In the human-computer interaction (HCI) domain there has been attempts to develop a technical language for modeling interaction [13] followed by efforts to ontologically generalize this view in other domains, e.g., [14], but the issue is clearly more general and of wider application.

The problem here is that the physical notion of interaction does not generalize well. Indeed, we have not developed a suitable framework for interaction beyond that of the physical laws. Without such framework we cannot establish how and to which extent an object may influence the behavior of another object. Yet, to make ontology suitable to evaluate interaction across humans and self-robots, a robust ontology of interaction must be developed.

5. Problems ahead of us

In the novel “The Book of Days”, Robert Chambers tells about a sow and her piglets charged and tried for the murder of a small child in 1457 [15]. Indicting a pig for a crime seems ridiculous today since, we believe, animals lack awareness of their actions and of the outcomes (interestingly, the book reports that “The sow was found guilty and condemned to death; but the pigs were acquitted on account of their youth, the bad example of their mother, and the absence of direct proof as to their having been concerned in the eating of the child.”[3]). A trial in which an animal is charged of crime is today unthinkable for many people, and the same people would be willing to wink at the substitution of a robot for the animal.

From the point of view of modern animal studies as well as of robotics, the conclusion vary depending on the animal species, the robot system, and the theory of consciousness one takes [16, 17]. The problem of founding and running a social and equal system that can comprise humans and self-robots remains widely open: we lack the basic principles about how to understand and organize such systems, and to measure their ethical status.

The previous observation triggers a series of topics, from issues rooted in cognitive and educational sciences to socio-technical organization and individual responsibility. Leaving these apart, we can imagine a classification of conditions for which social systems cannot even develop or survive beyond contingent or fortuitous circumstances. Among the foreseeable cases we can also imagine social systems that would very likely arise as an evolution of our existing systems. There is a need to study these conditions and how they relate to each other. On the ontological side, we should develop dedicated methodologies to understand the parameters to evaluate reliable interactions across different agent types, and even sustainable full fledged social systems.

Finally, most of the scenarios that we may want to elaborate about humans and self-robots are not completely new in our world. There are plenty of interactions between humans and other non-human animals which already give the gist of the ontological problems we are called to study. Nonetheless, there are two substantial differences between non-human animals and self-robots that push the problem at a different level. First, in modeling interactions humans have always taken the anthropological viewpoint. We have attributed to animals a worldview which apes our own. This move is not possible now. Second, human experience is limited to interactions with biological agents. For sure we can start from what we know about our mixed societies where human and non-human animals cohabit (with human consent or not) but we have to move beyond this if we want to be ready to develop a general social system or, at least, to reliably interact with self-robots.

[3]http://www.thebookofdays.com/months/jan/17.htm

References

[1] B. Siciliano and O. Khatib, *Springer handbook of robotics*, Springer, 2016.

[2] M. Ben-Ari and F. Mondada, *Elements of robotics*, Springer International Publishing, 2018.

[3] K. Rajan and A. Saffiotti, Towards a science of integrated AI and Robotics, *Artificial Intelligence* **247** (2017), 1–9. doi:10.1016/j.artint.2017.03.003.

[4] R. Mizoguchi, Y. Kitamura and S. Borgo, A unifying definition for artifact and biological functions, *Applied Ontology* **11**(2) (2016), 129–154. doi:10.3233/AO-160165.

[5] D. Paulius and Y. Sun, A Survey of Knowledge Representation in Service Robotics, *Robotics and Autonomous Systems* **118** (2019), 13–30. doi:10.1016/j.robot.2019.03.005.

[6] A. Olivares-Alarcos, D. Beßler, A. Khamis, P. Goncalves, M.K. Habib, J. Bermejo, M.E. Barreto, M. Diab, J. Rosell, J. Quintas, J. Olszewska, H. Nakawala, E. Pignaton, A. Gyrard, S. Borgo, G. Alenya, M. Beetz and H. Li, A Review and Comparison of Ontology-based Approaches to Robot Autonomy, *The Knowledge Engineering Review* (to appear).

[7] *IEEE Standard Ontologies for Robotics and Automation*, IEEE Stdandard 1872-2015, 10 April 2015.

[8] D. Lewis, *Convention - A Philosophical Study*, Blackwell Publishers, Bodmin, Cornwall (Great Britain), 1969.

[9] H. Ekbia and B. Nardi, Heteromation and its (dis) contents: The invisible division of labor between humans and machines, *First Monday* **19**(6) (2014). doi:10.5210/fm.v19i6.5331.

[10] S. Borgo and E. Blanzieri, Trait-based Culture and its Organization: Developing a Culture Enabler for Artificial Agents, in: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 333–338. doi:10.1109/IROS.2018.8593369.

[11] E.A.O. Diallo, A. Sugiyama and T. Sugawara, Coordinated behavior of cooperative agents using deep reinforcement learning, *Neurocomputing* (2019). doi:10.1016/j.neucom.2018.08.094.

[12] S. Borgo and C. Masolo, Foundational Choices in DOLCE, in: *Handbook on Ontologies*, 2nd edn, S. Staab and R. Studer, eds, Springer Verlag, 2009, pp. 361–381. doi:10.1007/978-3-540-92673-3_16.

[13] L.-E. Janlert and E. Stolterman, The Meaning of Interactivity—Some Proposals for Definitions and Measures, *Human–Computer Interaction* **32**(3) (2017), 103–138. doi:10.1080/07370024.2016.1226139.

[14] S. Borgo, An ontological view of components and interactions in behaviorally adaptive systems, *Journal of Integrated Design & Process Science* (2019), 1–19 (pre-press). doi:10.3233/JID190013.

[15] R. Chambers, *The book of days: a miscellany of popular antiquities in connection with the calendar, including anecdote, biography, & history, curiosities of literature and oddities of human life and character*, W. & R. Chambers, 1863.

[16] D. Gamez, Progress in machine consciousness, *Consciousness and Cognition* **17**(3) (2008), 887–910. doi:10.1016/j.concog.2007.04.005.

[17] J. Tani, *Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena*, Oxford University Press, 2016.