Looking for a metaphor of alternative autonomy to escape the being-machine duopoly in HRI

Christoph Salge and Marcus M. Scheunemann

University of Hertfordshire, Hatfield, AL10 9AB, UK

The robots of the future do not yet exist. Consequently, we *all* lack first hand experience, and must make do with stories and metaphors to think about them. Unsurprisingly, we often turn to fiction, and here we posit that fiction about robots has some common shortcomings, which in turn make it difficult to think about, write about and conduct HRI research—specifically HRI research dealing with the autonomous behavior generation for robots. Hence we are hoping to find a better metaphor.

The problem, in more detail, is that robots in fiction – and consequently in most of our thought landscape – exist mostly in a *being-machine duopoly* or further constrained, in a human-machine duopoly: They are either perceived as machines or humans—with little in between. Either they lack all aspects of living things, being seen as mere collection of rules and circuitry—some accounts even going so far as positing that they should not even be called AI. Then there are other robots, usually only found in fiction, that are basically humans in all but origin. They behave like humans, and will immediately be imbued with motivations, creativity, a will to life, and a right to freedom. This is likely due to the fact that we do have no real examples of anything existing in between. Living beings have a whole collection of aspects that make them living, and machines do not. It is hard to imagine that something that possesses some of those aspects does not possess all of them. We see this play out in fiction, such as books and movies, numerous times, where AIs or robots fall into one of two camps. One illustrative example is the treatment of robots in the Fallout game series [1]. But the question is: is there a middle ground, or even more radical, can we think of robots that even leave the linear interpolation between those two concepts?

We are particularly interested in this question, because a lot of our research deals with questions about robot's autonomously generating their behavior [3,4,5]. We care about equipping robots with intrinsic motivations, those drives that are essential to agency itself. In nature, those drives link back in their distal cause to constitutive autonomy and the preservation of each organism's precarious existence. These concepts can be, philosophically, linked to ideas such as Varela's autopoiesis [2]. In many of our works we spend quite some time talking about the idea of autonomy. Not just autonomy that goes from a scale of remote control to picking their own actions, but Autonomy as it relates to genuine goal ownership and meaningful actions as related to an agent's existence. On the other hand, we do have to be careful with our claims. Putting intrinsic motivations on a robot makes them by default extrinsic motivations. When we talk about computational models of intrinsic motivations we do refer to something that mimics those mechanisms that might arise from evolution and true

Autonomy, but we need to acknowledge that those are in the end just rules we put there, that somehow approximate those phenomena. We do this, because in some cases the behavior generated is robust and can be applied to a wide range of scenarios and morphologies, giving results that can often best be described as interesting, because of the lack of a good evaluation function. We recently also had some indications that putting intrinsic motivations on a robot might make them appear to a human interaction partner more as a genuine social other – increasing their perception of the robot's warmth [5]. But there are a lot of conceptual issues related to this research. If the above paragraph sounded a bit convoluted, and difficult, it is because this might be the most accessible way to talk about it—which is not good. Explanations of what this level of autonomy is are difficult. Measures for how it should look like are even harder to come by. When we explain what is happening, there is often the desire to put this either in more mechanical terms, given a too rigid and mechanical explanation, or to put it into too meaningful terms for living beings.

What we want is a proper metaphor, and ideally some good language to go along with it, that allows us to talk about this alternative autonomy. We do believe that the concept of autonomy, more fundamentally understood, is a key component to understanding what we are missing here, as it allows us to pinpoint a part of the conceptual landscape that looks very murky, and is massively under-explored. It also charts a path that we might have to travel on, because the robots of the future seem to be in possession of a lot of autonomy, and unlike fiction, HRI can probably not simply jump from no autonomy and machine-like beings to full autonomy and living robots without any steps in between.

References

1. Götter, C., Salge, C.: From Deep Blue to Blade Runner – The Portrayal of Artificial Intelligence in the Fallout Game Series. Paidia (2017)
2. Maturana, H.R., Varela, F.J.: Autopoiesis and Cognition: The Realization of the Living, vol. 42. Springer Science & Business Media (1991). https://doi.org/10.1007/978-94-009-8947-4
3. Scheunemann, M.M., Cuijpers, R.H., Salge, C.: Warmth and Competence to Predict Human Preference of Robot Behavior in Physical Human-Robot Interaction. In: 29th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp. 1340–1347. IEEE (2020). https://doi.org/10.1109/RO-MAN47096.2020.9223478
4. Scheunemann, M.M., Salge, C., Dautenhahn, K.: Intrinsically Motivated Autonomy in Human-Robot Interaction: Human Perception of Predictive Information in Robots. In: Towards Autonomous Robotic Systems. pp. 325–337. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-23807-0_27
5. Scheunemann, M.M., Salge, C., Polani, D., Dautenhahn, K.: Human Perception of Intrinsically Motivated Autonomy in Human-Robot Interaction (2020), https://arxiv.org/abs/2002.05936