Expectation and anticipatory systems for AI

Harko Verhagen and Teresa Cerratto-Pargman,

Dept of Computer and Systems Sciences, Stockholm University, Sweden

Extended abstract

Following the definition in Rosen (2012) of anticipatory systems as systems that adapt their behavior using models of the overall system of which they are part, more in particular predictions of future states as well as his statement that all living systems are anticipatory systems we cannot conclude that the reverse is true – that all anticipatory systems are living systems. Indeed, simplistic models have long been used to implement some level of agency in control systems, for instance a thermostat. The development of artificial intelligence (AI) has enabled the implementation of higher levels of agency.

The agency of AI artefacts is *a priori* given and limited both explicitly and implicitly limited at the time of their creation, by the designer. Explicitly in the way such agency can function in and interact with the overall system. Implicitly in the way agency is embedded in and imbued with the designer's own values and norms. However, AI artefacts can – given that there is a provision for learning after the design phase – overcome the first, explicit limitation. The second, implicit limitation that comprises designer's values and norms is much harder to surpass. Theoretically, this issue has been described in for instance the work on norm-autonomous agents (Chopra et al., 2018) where the highest level of agency is defined as the freedom for the artefact to define its own norms and values (Verhagen, 2000).

In general most AI systems only evaluate the physical aspects of their environment in an anticipatory way. At the norm-autonomous level we face challenges as we need to engage with modelling of and reasoning about social and thus moral and ethical aspects of the overall system in which the artefact is to function. Such reasoning broadens the artificial anticipatory systems from been physically embedded to become socially and cultural historically embedded systems. It extends beyond teleological reasoning to normative, ethical and moral reasoning in the artificial anticipatory systems, if they are to function well in hybrid social systems comprising human and artificial agents.

Some implementations of such systems do exist in for instance social simulation models while in physical artefacts such as autonomous vehicles the anticipatory model is mostly limited to the physical world. While there is discussion on ethical and legal aspects (for instance using the well-known Trolley problem example) still little is done regarding the social aspects inherent to the interaction between the artificial and human anticipatory systems. Some of the problems with autonomous vehicles are related to these issues.

Against this background, we suggest the concept of expectation can be seen as a central concept to move the discussion about anticipatory systems forward. While the concept of anticipation needs to have a model of changes in the world due to actions of entities in it, such model can be used to create expectations that may or may not come true – the basis for learning of action-consequences and sequences. More in particular, the concept of expectation is instrumental in taking into account the social context by modeling the norms and values active in that context, allowing for normative expectations to be part of the anticipatory system next to the expectations based on learned action-consequence sequences. This will enable the artificial agent to participate in social practices (Reckwitz, 2002).

This creates a set of questions to explore. How do designers engage with expectation and social practices in their practices? More specifically how do they take account of the social, ethical and moral dimensions of living systems when designing the agency of AI artifacts? And how do designers incorporate the expectations the artefact they are developing may cause in the users?

How do designers take account, make explicit of their own social, ethical, moral values and norms in the design of AI artifacts? And what is the role played by designers' values and norms in the design of agency of Ai artifacts? What role plays designer's awareness of "the Other" (e.g., in other cultures) in the design of AI artifacts? And finally, how can designers be reflective on their own values, norms and understandings of otherness in the design of AI artifacts?

References

A. Chopra, L. van der Torre, H. Verhagen, S. Villata (Eds.). Handbook of Normative Multiagent Systems. College Publications, 2018.

Reckwitz, A. Toward a theory of social practices: A development in culturalist theorizing. European journal of social theory, 5(2), pp.243-263, 2002.

R. Rosen: Anticipatory Systems - Philosophical, Mathematical and Methodological Foundations. Pergamon Press, London, 1985; 2nd Ed: Springer, 2012.

H. Verhagen. Norm Autonomous Agents (PhD thesis). DSV technical report 00-004, 2000.