Editorial: Agency in Natural and Artificial Systems

Marieke Rohde & Takashi Ikegami

The 1980s and 1990s saw a re-orientation of the field of artificial intelligence (AI) away from disembodied software, a domain where AI research continues predominantly as software engineering, and toward situated and embodied robotics. Crucially, the publication of Braitenberg's (1984) *Vehicles* and, later on, the behavior-based robotics movement (e.g., Brooks, 1995) catalyzed this emphasis on *agency* rather than abstract reason, by demonstrating that circuits embedded in closed sensorimotor loops can exhibit adaptive capacities that exceed those of disembodied software.

In the meantime, both artificial life approaches (e.g., Brooks', 1995, subsumption architectures; evolutionary robotics: see Harvey, Di Paolo, Wood, Quinn, & Tuci, 2005 and Nolfi & Floreano, 2000), that focus on sensorimotor circuitry, and more conventional approaches (e.g., commercial robots, such as Sony's AIBO or Honda's ASIMO), that rely on explicit design and planning, have succeeded in producing life-like behaviors. Such artifacts appeal to us, attract our interest, and evoke our sympathy. In a 21st century variant of the Turing test (Turing, 1950) that probes our capacity to distinguish an artificial system's embodied agency from natural agency, not its verbal communication skills, such artifacts pass with flying colors.

Yet, something still appears to be missing in current robotics. Criticism has been phrased that such machines merely look *as if* they act, *as if* they are smart, *as if* they have intentions (e.g., Di Paolo, 2003) or that Turing-test style approaches that focus on the surface behavior need to be complemented with the study of behavior generating mechanisms to avoid falling victim to a farce (Rohde & Stewart, 2008). The gist of such criticisms is that, unlike humans or animals, current robots do not really care about the success of their actions. Moreover, the situations in which robot behavior breaks down reveal a lack of even understanding the goal of their actions in the first place. Arguably, a robot that moves but does not understand what its actions should achieve is not really an agent in a strong sense, it only acts *as if*.

Biology shows us that it is possible for a physical system to act according to inherent goals, something we are still missing out in current robotics. The question is whether what is missing is another mechanism to be added to current robot architectures, or whether our modeling approaches have to be rethought in a more fundamental way. *How can we build machines that possess stronger forms of agency?* The goal of this special issue is to advance on this fundamental problem, identifying new principles for the synthesis of agency. The contributions share in common that they identify *complex processes of dynamical selforganization on several levels* as possible routes toward obtaining genuine agency.

The problem had been previously addressed in a three-day workshop on Agency in Natural and Artificial Systems in Kyoto in 2008, organized by the editors of this issue, together with Jun Tani and Ezequiel Di Paolo, where the idea to compile this special issue was born. Many of the ideas underlying the contributions were generated in the inspiring atmosphere of the Kansai seminar house in the hills surrounding Kyoto. Authors include both participants of the workshop and external contributors, and all submissions underwent a process of double blind review.

Besides the underlying emphasis on dynamical self-organization on different levels, a number of motifs recur in this special issue, sometimes more in the foreground and sometimes more subtly in the background. These are the problem of *agency detection, the*

observer, and the generative mechanisms, the relation between *self-maintenance, boundary construction, and acting in an environment* and the issue of *inherent spatiotemporal constraints*.

There is a conceptual observer/frame problem about detecting agency in a system, which can only be done by another agent: our claim is that we have not generated real agency yet —but how will we know we have obtained our goal? In answering this question, there is a tension between criteria that focus on the generative mechanisms and those that focus on the appeal of surface behavior. Criteria that focus on the mechanism run into danger of neglecting the inter-subjective nature of the problem of agency detection, whereas behavioral Turing-test style approaches run into danger of favoring superficial imitation over genuine instantiation (Rohde & Stewart, 2008). This problem is most explicitly addressed by Aucouturier and Ikegami, who propose a dynamical Turing test. Also, Tani's neurorobotics study explicitly pays heed to the problem of the observer in artificial agent research by ongoing reference to the phenomenology of self.

In natural agents, the processes of self-construction and those that generate behavior are functionally and mechanistically intertwined. For instance, human cognitive neuroscience (e.g., Ramachandran, 1998) shows that our body boundary is dynamically constructed through an interplay of sensorimotor behavior and neural change. Autopoietic theory even states that emergent self-construction from a network of local interactions is one of the defining properties of life and cognition (Maturana & Varela, 1980). In most artificial agent research, the body is externally defined and rigid, or otherwise body image plasticity is implemented as a separate dedicated and centrally controlled mechanism, which is not biologically plausible. On the other hand, proto-cell models (e.g., Varela, Maturana, & Uribe, 1974) have no clear distinction between behavior generation and self-construction at all, which means that behavior is always immediately tied to, and limited by, viability constraints, which is not what we find in more complex organisms either. The attempt to integrate and relate these two levels is most explicitly made by Barandiaran, Di Paolo, and Rohde (conceptually) and by Egbert and Di Paolo (in a model).

A third recurring theme is the issue of (spatio-) temporal embeddedness: real world physics imposes constraints on natural agents and their interactions with the environment. Turn-taking in dyadic interactions (Tani) and other socially coordinated processes (De Jaegher & Froese), as well as interaction with the inanimate environment through phases of coupling and decoupling (Aucouturier & Ikegami) are important properties of natural agents that require an appropriate spatio-temporal structuring of behavior. Barandiaran et al. argue that an agent's own spatial and temporal meaning domain emerges from the inherent dynamics of behavior and interaction and that such subjective space-time is not necessarily Euclidean. Usually, in computational models, the experimenter arbitrarily defines a time-step for the agent controller off-line, such that it can exhibit useful spatio-temporal behavior when embedded in closed-loop interaction. This choice is not an inherent physical constraint; it is not plastic, it is outside the agent's control or access and independent of ongoing sensorimotor behavior. The complex inherent dynamics of the neural robot controllers presented by Aucouturier and Ikegami (itinerating chaos) and by Tani (self-organizing criticality) manifest in coordinated behavior in the closed loop that is difficult to mimic if programmed explicitly. Such coordination emerging from an interplay between internal dynamics and agent-environment-interaction dynamics is a first step towards inherent temporality of agency.

In the following, the five contributions are briefly introduced, in the order in which they are published.

Barandiaran et al. propose a definition of agency that is based on the requirements of

individuality, interactional asymmetry, and normativity. The definition aims to capture what is missing from most current definitions of agency and is rooted in autopoietic theory (Maturana & Varela, 1980) and related approaches to biological autonomy. The requirements of normativity and individuality relate to these ideas. The authors propose that agency additionally requires a system to autonomously regulate its interaction with the environment in order to obtain a goal, even if such regulation only manifests as point-wise modulation of ongoing behavior (interactional asymmetry). They evaluate their proposal in terms of spatio-temporality of behavior and apply it to existing models, as well as those envisioned for the future.

Egbert and Di Paolo propose a model that integrates autopoiesis and spatio-temporal behavior with a purpose. Their model is based on an artificial chemistry and involves the self-maintenance of a metabolic network inside a self-organizing membrane, in the spirit of models of autopoiesis (Varela et al., 1974). However, unlike most proto-cell models that focus on self-maintenance, the system also performs chemotactic behavior in the simulated environment. Both motor behavior and metabolistic self-maintenance emerge from the same set of reactions in the underlying artificial chemistry, even if they are functionally distinct. The authors analyze the model in the light of the question of how mechanism and function intermingle or dissociate both at the level of behavior and at the level of physico-chemical processes. The model strongly links to the definition and requirements proposed in Barandiaran et al.

Aucouturier and Ikegami compare two approaches to modeling agency, that is, an explicit constraint logic programming approach that imitates spatio-temporal properties of natural agents, and a chaotic dynamical neural network approach, in which such characteristics are naturally afforded. They analyze the dynamics of alternating phases of coupling and decoupling with the environment in terms of mutual information circulation. In a dancing robot, itinerating chaotic attractors in neural network controllers can bring about such alternation between coupling and decoupling with the rhythm of music that are reminiscent of the dynamics of humans dancing for enjoyment. The authors emphasize the behavioral aspect of agency (in terms of agency detection by an external observer) and propose a “dynamical Turing test” to establish agency. In this sense, they complement the mechanismbased approach taken by Barandiaran et al. and Egbert and Di Paolo.

Tani presents a series of neuro-robotic experiments that are inspired by classical phenomenology of subjectivity (i.e., being a self). In a similar spirit as Aucouturier and Ikegami, Tani presents neural controllers exhibiting dynamics that naturally afford properties that are characteristic of natural agency and individuality. The phenomena treated in succession are automation of behavioral routines and their breakdown, imitation and turn-taking in social interaction and reflexive self-referentiality. His account is rooted in first person experience of agency, not in third person agency detection or study of the generative mechanism. Tani identifies the phenomenon of selforganizing criticality of sensorimotor dynamics with the occurrence of spontaneous switching or breakdown in the different behavioral domains (for instance, when spontaneously initiating turn-taking in social interaction). The observation that dynamically critical events are characteristic of agency relates to Barandiaran et al.’s discussion of interactional asymmetry as punctuated modulation of ongoing behavior as a hallmark of agency.

De Jaegher and Froese, with a primarily conceptual contribution, assess to what extent processes of dynamical self-organizing between individuals (i.e., in interaction) can be relevant for the study and generation of agency. They point out that the dynamical complexity that results from multi-agent interaction can be as powerful as, or even more powerful than, the dynamical complexity resulting from sensorimotor self-organization in a closed-loop. Examples are given for how the characteristics of natural agency (such as forming intentions,

performing a goal-directed action, acquiring a skill) can be either catalyzed or disrupted by interactional dynamics between individuals. The authors argue that, when studying the dynamics of participatory sense-making between individuals, modelers will be able to address questions of higher complexity than when studying individual agents on their own. Their contribution can be seen as complementary to the work presented, which, with the exception of Tani's neurorobotic experiment on imitation learning in human robot interaction, focuses on the individual agent and its interaction with a world without other agents.

Many of the big questions that came up during the 2008 workshop in Kyoto are analyzed, discussed, and sometimes even answered in the contributions to this special issue. The special issue has a strong conceptual emphasis—the authors do not provide simple models or general recipes to solve our problems all at once. However, they give concrete proposals on how to advance with the problem of agency step by step. In order to generate agency, we first have to understand and define how we, as external observers, assess, define, and experience agency, both on the basis of surface behavior and on the basis of generative mechanisms. Robot controllers should have rich internal dynamics that intrinsically afford the dynamical patterns of spatio-temporally embedded behavior that make us perceive them as agents. The agent's body and the spatio-temporal properties of its interactions should not be externally defined, but should emerge from, or be dynamically self-constructed in, interaction. Dynamical interaction should not be restricted to an inanimate spatial environment, but should include other agents to attain higher levels of dynamical and cognitive complexity. These concrete proposals point out what it is that is missing in current artificial agent research and enable us to conceive how we can go beyond the external specification of purpose, the teleonomy that currently marks both traditional and behavior-based approaches to agency.

Acknowledgments

We are very grateful to the contributors to have put so much work in this ambitious and versatile special issue. We are even more grateful to the reviewers, who have been very thorough and reliable, even when we confronted them with the most inappropriate demands and deadlines. Marieke Rohde wishes to acknowledge the support of the HFSP and the JSPS (short-term fellowship) for funding her work.

References

- Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.
- Brooks, R. (1995). Intelligence without reason. In L. Steels & R. Brooks (Eds.), *The artificial life route to artificial intelligence: Building embodied, situated agents*. Hillsdale, NJ: Lawrence Erlbaum.
- Di Paolo, E. A. (2003). Organismically inspired robotics: Homeostatic adaptation and natural teleology beyond the closed sensorimotor loop. In K. Murase & T. Asakura (Eds.), *Dynamical systems approach to embodiment and sociality* (pp. 19–42). Adelaide, Australia: Advanced Knowledge International.
- Harvey, I., Di Paolo, E., Wood, R., Quinn, M., & Tuci, E. A. (2005). Evolutionary robotics: A new scientific tool for studying cognition. *Artificial Life*, *11*(1–2), 79–98.
- Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Boston, MA: D. Reidel.
-Nolfi, S., & Floreano, D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. Cambridge, MA: MIT Press.
- Ramachandran, V. S. (1998). Consciousness and body image: lessons from phantom limbs, Capgras syndrome and pain asymbolia. *Philosophical Transactions of the Royal Society of London B Biological Science*, *353*(1377), 1851–1859.
- Rohde, M., & Stewart, J. (2008). Ascriptional and 'genuine' autonomy. *BioSystems*, *91*(2), 424–433. [Special issue on modelling autonomy.]
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *49*, 433–460.
- Varela, F., Maturana, H., & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems*, *5*, 187–196.