Rapaport, William J. "Meinongian Semantics and Artificial Intelligence." *Humana.Mente: Journal of Philosophical Studies* 25 (2013): 25–52. Available at http://www.cse.buffalo.edu/~rapaport/Papers/rapaport-humanamente.pdf.

Rapaport, William J. "On the Relation of Computing to the World." 2015 IACAP Covey Award Keynote Address. Available at http://www.cse.buffalo.edu/~rapaport/Papers/covey.pdf.

Rapaport, William J., and Michael W. Kibby. "Contextual Vocabulary Acquisition as Computational Philosophy and as Philosophical Computation." *Journal of Experimental and Theoretical Artificial Intelligence* 19, no. 1 (2007): 1–17. Available at http://www.cse.buffalo.edu/~rapaport/Papers/cva-jetai.pdf.

Rapaport, William J., and Michael W. Kibby. "Contextual Vocabulary Acquisition: From Algorithm to Curriculum." In *Castañeda and His Guises: Essays on the Work of Hector-Neri Castañeda*, edited by Adriano Palma, 107–50. Berlin: Walter de Gruyter, 2014. Available at http://www.cse.buffalo.edu/~rapaport/Papers/reading4HNC.pdf.

Rapaport, William J., and Stuart C. Shapiro. "Cognition and Fiction." In *Deixis in Narrative: A Cognitive Science Perspective*, edited by Judith Felson Duchan, Gail A. Bruder, and Lynne E. Hewitt, 107–28. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995. Available at http://www.cse.buffalo.edu/~rapaport/Papers/rapaport.shapiro.95.cogandfict.pdf.

Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (1980): 417–57.

Searle, John R. "The Myth of the Computer." *New York Review of Books*, April 29, 1982, 3–6. Cf. correspondence, same journal, June 24, 1982, pp. 56–57.

Searle, John R. "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): 21–37. Reprinted in slightly revised form in Searle, *The Rediscovery of the Mind*. Cambridge, MA: The MIT Press, 1992, Ch. 9.

Searle, John R. "The Failures of Computationalism." *Think* 2 (1993): 68–71. Tilburg University Institute for Language Technology and Artificial Intelligence, Tilburg, The Netherlands. Available at http://users.ecs.soton.ac.uk/harnad/Papers/Harnad/harnad93.symb.anal.net.searle.html.

Shapiro, Stuart C., and William J. Rapaport. "SNePS Considered as a Fully Intensional Propositional Semantic Network." In *The Knowledge Frontier: Essays in the Representation of Knowledge*, edited by Nick Cercone and Gordon McCalla, 262–315. New York: Springer-Verlag, 1987. Available at http://www.cse.buffalo.edu/~rapaport/Papers/shapiro.rapaport.87.pdf.

Shapiro, Stuart C., and William J. Rapaport. "Models and Minds: Knowledge Representation for Natural-Language Competence." In *Philosophy and AI: Essays at the Interface*, edited by Robert Cummins and John Pollock, 215–59. Cambridge, MA: The MIT Press, 1991. Available at http://www.cse.buffalo.edu/~rapaport/Papers/mandm.tr.pdf.

Shapiro, Stuart C., and William J. Rapaport. "The "SNePS Family." *Computers and Mathematics with Applications* 23 (1992): 243–75. Reprinted in *Semantic Networks in Artificial Intelligence*, edited by Fritz Lehmann, 243–75. Oxford: Pergamon Press, 1992. Available at http://www.sciencedirect.com/science/article/pii/0898122192901436.

Shapiro, Stuart C., and William J. Rapaport. "An Introduction to a Computational Reader of Narratives." In *Deixis in Narrative: A Cognitive Science Perspective*, edited by Judith Felson Duchan, Gail A. Bruder, and Lynne E. Hewitt, 79–105. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995. Available at http://www.cse.buffalo.edu/~rapaport/Papers/shapiro.rapaport.95.pdf.

Smith, Brian Cantwell. "The Correspondence Continuum." *Technical Report CSLI-87-71*. Center for the Study of Language & Information, Stanford, CA, 1987.

Wilkins, David P. "Expanding the Traditional Category of Deictic Elements: Interjections as Deictics." In *Deixis in Narrative: A Cognitive Science Perspective*, edited by Judith Felson Duchan, Gail A. Bruder, and Lynne E. Hewitt, 359–86. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995. Available at http://www.cse.buffalo.edu/~rapaport/dc.html.

From Biological to Synthetic Neurorobotics Approaches to Understanding the Structure Essential to Consciousness (Part 3)

Jeff White
INDEPENDENT SCHOLAR

Jun Tani
OKINAWA INSTITUTE OF SCIENCE AND TECHNOLOGY (OIST)
TANI1216JP@GMAIL.COM

ABSTRACT

This third paper locates the synthetic neurorobotics research reviewed in the second paper in terms of themes introduced in the first paper. It begins with biological non-reductionism as understood by Searle. It emphasizes the role of synthetic neurorobotics studies in accessing the dynamic structure essential to consciousness with a focus on system criticality and self. It develops a distinction between simulated and formal consciousness based on this emphasis, reviews Tani's and colleagues' work in light of this distinction, and ends by forecasting the increasing importance of synthetic neurorobotics studies for cognitive science and philosophy of mind going forward, finally in regards to most- and myth-consciousness.

1. KNOCKING ON THE DOOR OF THE CHINESE ROOM

> Prediction is made possible by adaptive mechanisms that are supported by learning rules that either apply across generations (evolutionary adaptation) or within the lifetime of the organism. As a result, organisms can deal with a future occurrence of the same or similar situations more effectively. This is the fundamental organization principle of any adaptive system.
>
> – Buszaki, Pyerache, and Kubie[1]

This series began with Boltuc's "Is anyone home?" question,[2] responding with a sketch of an agent proactively invested in integrating past with present in order to achieve an optimal future. Contrary to Boltuc's naturalistic nonreductionism recommending that a "projector" of consciousness be first resolved in order to engineer similar in an artificial agent, we rejected the notion that consciousness can be isolated to any loci of activity, arguing that formal articulation of essential dynamics in synthetic neurorobots opens a view on the problem of consciousness that is not available to biological inquiry, alone. That first paper concluded with an introduction to, and the second paper continued with a detailed review of, two decades of research by Jun Tani and colleagues accounting for self, free will and consciousness in neurorobots within the predictive coding framework and according with the free energy principle. Central to this review was the notion of system criticality, with which the continuous perceptual stream is segmented and episodes rendered objects for later recall and recomposition, and which remains central to the current paper, as well.

The present paper proposes the notion of "formal" consciousness to distinguish systems which aim to resolve the source of subjectivity in system criticality from work aiming for other ends, "simulations" and "reasonable approximations" of human consciousness for example intent on passing a Turing test without regard for first person phenomena. This section briefly locates this position in the contemporary context. The following section reviews Tani and colleagues' neurorobotics research aimed at understanding consciousness with a focus on the notion of criticality, and how incoherence and the breakdown of established and anticipated patterns opens a privileged view on the emergent self and consciousness thereof. The third section delineates formal consciousness in terms of three necessary factors present in Tani and colleagues' work yet absent in others, and the fourth section forecasts that synthetic neurorobotics will play an increasingly central role in consciousness studies going forward.

At the turn of the last century, John Searle found the problem of consciousness the most pressing open to biological inquiry and explanation. He faulted assumptions that the rejection of either dualism or materialism compelled the adoption of the other, and championed biological naturalism as an alternative. He wrote:

> We know enough about how the world works to know that consciousness is a biological phenomenon caused by brain processes and realized in the structure of the brain. It is irreducible not because it is ineffable or mysterious, but because it has a first-person ontology and therefore cannot be reduced to phenomena with a third-person ontology.[3]

This distinction between first and third person ontologies helps to frame the hard problem of consciousness, which for students of artificial consciousness is perhaps most clear in Searle's distinction between semantics and syntax. A machine performs syntactical operations while human beings (conscious) do something more, they *understand*, a point originally illustrated in Searle's famous Chinese Room thought experiment.[4]

Searle's Chinese room is an argument against reductive physicalism, and equally against the notion that consciousness is software running on hardware as in a modern digital computer. It illustrates that there is something missing in the mere exchange of symbols at which computers are so proficient, and casts doubt on how a "Turing test" might confirm consciousness. After all, the "imitation game" was not originally conceived of as a test for consciousness, but rather as a test for the ascription of intelligence. The question was "Can machines think?" and more importantly, can thinking machines be indiscernible from human beings in doing so?[5]

On Searle's understanding, computational hardware pushes symbols according to a program.[6] Computers do not evolve in material interaction with a pervasive natural world, as do human beings, and do not become conscious through this interaction. They are not autonomous; they are programmed. The best that such a machine can do is to simulate, or approximate, consciousness, and they do so by explicit design. Accordingly, simulated consciousness is not consciousness on Searle's account, but he did not bar the door on artificial consciousness, either. Rather, he pointed to where the key to such may be found. He wrote that "understanding the nature of consciousness crucially requires understanding how brain processes cause and realize consciousness"[7] and that conscious artifacts may be designed which "duplicate, and not merely simulate, the causal powers that [biological] brains have"[8] once such an understanding is achieved.

As a positive research program, Searle recommended correlating neurobiological activity with conscious phenomena, checking for causal relationships, and developing laws formalizing these relationships.[9] He identified two ways forward in this industry, the "building blocks"[10] and "unified field"[11] approaches, but dismissed the former because "The production of any state of consciousness at all by the brain is the production of a unified consciousness."[12] At that time, he pointed to Llinas et al. and Tononi, Edelman, and Sporns as examples of unified field friendly approaches, involving the top-down integration of system wide information within the thalamocortical region.[13]

Since that time, Tononi and colleagues have developed the Integrated Information Theory (IIT). According to the IIT, consciousness does not require "contact with the external world" but rather "as long as a system has the right internal architecture and forms a complex capable of discriminating a large number of internal states, it would be highly conscious."[14] The "integration" of IIT implies that such a system be unified and seek to maintain this unity in the face of disintegrative change, with each part of the system able to be affected by any other part of the system as measured by the irreducibility of its intrinsic cause-effect structure. A biological brain exemplifies maximal intrinsic irreducibility as a cause-effect structure with definite borders and highly integrated information.[15] Other systems are irreducible, for example two men in conversation, but are not maximally irreducible intrinsically as they are not fully integrated. So understood, "consciousness is not an all-or-none property," but it is not open to piecemeal assembly either, rather increasing with "a system's repertoire of discriminable states."[16] At the minimal level, a "minimally conscious system" distinguishes between just two "concepts"[17] such that "even a binary photo-diode . . . enjoys exactly 1 bit of consciousness"[18] and systems increase from there with their discriminable states.

In conjunction with quantity of consciousness, quality of consciousness derives from the structure affording it, and the IIT leaves it to engineers to delimit the contents of artificial consciousness by "appropriately structuring" an agent's "effective information matrix."[19] As for determining which structures deliver which qualities, Tononi and colleagues also suggest that inquiry begin with biological models, with this understanding first tested against personal and then extended to all human experience before duplication in artificial systems. In the end, the "IIT predicts that whatever the neural correlate of consciousness (NCC) turns out to be" it will be the locus of

integration over discriminable states which "may expand, shrink and even move within a given brain depending on various conditions."[20] Thus, the IIT continues in Searle's line of reasoning.

Contrast the view put forward by leading commercial roboticist Theodore Goertzel. Goertzel does not aim to duplicate but rather at a "reasonable approximation" of three persistent aspects of consciousness, "free will, reflective consciousness" and "phenomenal self." What is "important" for Goertzel is "to identify the patterns constituting a given phenomenon" and trace "the relationships between various qualities that these patterns are hypothesized to possess (experiential versus physical)," an approach reinforced by the observation that "from the point of view of studying brains, building AI systems or conducting our everyday lives, it is generally the patterns (and their subpatterns) that matter" with given phenomena "understood" as correlate activity patterns are identified.[21]

Goertzel's "patternism" is appealing. It is consistent with calls for the qualification of artificial systems by biological activity. Furthermore, the focal shift from neural loci to activity patterns coincides with advancing inquiry into biological substrates of consciousness, as current imaging technologies afford the establishment of functional correlations between networked neural dynamics in biological models and self-reports of various aspects of consciousness. In light of such advancing research for example, Searle's "already conscious" can be re-assessed in terms of the resting state "default" network based in the ventromedial prefrontal cortex and the posterior cingulate cortex.[22] Heine et al. affirm the promise in interpreting the conditions of non-communicating subjects through the lens of such activity patterns, a lens that may be repurposed in the evaluation of artificial agents of appropriate architectures which also may not self-report and indeed may not interact with the external world as we know it.[23] Such patterns can be then mapped onto Goertzel's freewill, reflective consciousness and phenomenal self, underscoring the potential of this approach in evaluating non-biological systems in similar terms.

However, there remain doubts that consciousness is realized in duplicate activity patterns, alone. For example, Oizumi et al. characterize patterns of activity internal to the cognitive agent in terms of "shapes" in "concept" and "phenomenal space" exported as graphical representations, at the same time warning that "one needs to investigate not just ''what'' functions are being performed by a system, but also ''how'' they are performed within the system."[24] On the IIT, it is the integration over discernible system states that is essential to consciousness, with "strong" integrated systems autonomous as they act and react from internally composed states and goals.[25] On this account, pattern matching alone does not achieve the strong integration that IIT demands. For one, patterns are not necessarily "strongly" integrated, i.e., fully embodied and constrained by the possible futures that this embodiment affords, i.e., maximally irreducible intrinsically. Furthermore, without such strong integration, there is no experience. Accordingly, overt focus on patterns—"what"—exclusive of how (and why) they arise opens the door to "true" zombies exhibiting "input output behavior" approximating biological activity patterns "while lacking subjective experience" at the same time.[26]

In summary, Goertzel's "reasonable approximation" might open the door to the Chinese room, but as zombie patterns should be indiscernible from non-zombie patterns, what greets us may be a zombie. For the patternist, this may not be a problem. Goertzel's goal is passing a Turing Test for which a reasonable approximation may suffice. But, when it comes to confirmation of consciousness in an artifact, it clearly does not, as captured in the concern that we may build a system "behaviourally indistinguishable from us, and certainly capable of passing the Turing test" that remains a "perfect" zombie at the same time.[27]

In 2009, Jun Tani noted a similar limitation in existing examples of machine intelligence such as behavior-based robotics articulating sensory-motor reflex behaviors. On his assay, systems aimed at passing the Turing test "turn out to be just machines having stochastic state transition tables" and

> after a while, we may begin to feel that the robots with reflex behaviors are simply like steel balls in pinball machines, repeatedly bouncing against the pins until they finally disappear down the holes.[28]

Further, Tani asks,

> But what is wrong with these robots? Although they have neither complex skills for action nor complex concepts for conversation, such complexity issues may not be the main problem.[29]

Instead, Tani argues that "the problem originates from a fundamental lack of phenomenological constructs in those robotic agents" and that "[i]n particular, what is missing … [is] . . . the "subjectivity" that should direct their intentionality to project their own particular images on the outer objective world."[30] He goes on to suggest that subjectivity develops gradually through sensorimotor experience of an agent's direct interaction with the world.[31] As each robot is distinctly located in a shared space of action in terms of a shared objective world, each robot develops its own views as particular internal models that then enable it to anticipate and to interpret the outcomes of its actions, with moreover this shared metric space grounding a capacity to generalize these internal constructs in the communication with and interpretation of others similarly situated (see the second paper in this series for in-depth review).

Consider this issue in terms of identifying agency, as set out by Barandiaran, Di Paolo, and Rohde.[32] They consider that a necessary condition for agency is a system capable of defining its own identity as an individual, thus distinguishing itself from its surroundings including other agents. Of particular interest here is their view that the boundary of an individual is self-defined through interaction with the environment. Tani argues that the same dynamic grounds the emergence of subjectivity in the following way.[33]

Top-down anticipation may not correlate with perceived reality in many situations. When environmental interactions

proceed exactly as expected, behaviors can be generated smoothly and automatically. However, anticipation can sometimes be wrong, and the conflict that arises in such cases can make generating successive acts difficult. When environmental interactions cause the agent to shift spontaneously between opposite poles, from automaticity to conflict necessitating autonomy, the boundary between the subjective mind and the objective world fluctuates, and so the boundaries of self are realized. Here, Tani argues that the essential characteristics of this phenomenon are best understood in terms of traditional phenomenology, since phenomenologists have already investigated the first-personal characteristics of autonomous and authentic selves.[34] In the end, Tani expects that uncovering the mechanisms grounding autonomy will lead to understanding the dynamic structure essential to consciousness in terms consistent with those postulated by William James,[35] in terms of momentary selves in the stream of consciousness. The next section reviews Tani and colleagues' work in clarifying these mechanisms and the dynamics essential to self and consciousness that they reveal.

2. ANSWERING THE DOOR OF THE CHINESE ROOM

> Acts are owned as they adaptively assert the constitution of the agent. Thus, awareness for different aspects of agency experience, such as the initiation of action, the effort exerted in controlling it, or the achievement of the desired effect, can be accounted for by processes involved in maintaining the sensorimotor organization that enables these interactions with the world.
>
> – Buhrmann and Di Paolo[36]

How is consciousness to be assessed if not through a Turing test or via correlation with biological activity patterns? Paraphrasing Searle, approximations cannot be conscious. What about self-reports, then? "In neuroscience, the ability to report is usually considered as the gold standard for assessing the presence of consciousness."[37] Reporting on internal processes is *prima facie* evidence for the feeling of undergoing them. But again, this is no more a guarantee of consciousness than a Turing test, at once neglecting those systems unable to so report.

In the first paper, we made the case that computational models open consciousness to inspection where study of biological models alone cannot. We characterized these systems and their transitions in terms of predictive coding which aims at minimizing error by optimizing internal models guiding action, in biological models understood in terms of the "predictive brain."[38] In general terms, cognition manages transitions between situations by internalizing their dynamics, modeling their likelihoods, and preparing for them accordingly with the aim being the minimization of error in this process. Tani's thesis is that, where model and reality diverge and error is not minimal, consciousness arises in the effort of minimizing the difference by modifying the contextual state that the agent extends from the past in order to return to coherence with its situation. Before proceeding to show how Tani and colleagues are able to expose these dynamics and their relation to consciousness, a brief review of the free energy principle and its role in the emergence of the phenomenon of self is required. From this review, we will be in a position to better appreciate Tani's thesis on the emergence of self and consciousness, and its implication that the free energy principle, as with activity patterns and strong integration, cannot by themselves account for consciousness.

In the second paper, we reviewed Karl Friston's "free energy principle" by which an agent aims to minimize error (or "surprise") by maximizing the likelihood of its own predictive models. This approach extends natural processes and the energetics that characterize them into the sphere of cognitive systems consistent with other theses on the nature of cognition, from Helmholtz's unconscious inference to contemporary deep learning. Friston writes that "the time-average of free energy" "is simply called "action" in physics" and that "the free-energy principle is nothing more than principle of least action, applied to information theory."[39] "The free-energy principle simply gathers these ideas together and summarizes their imperative in terms of minimizing free energy (or surprise)" while also bringing "something else to the table . . . that action should also minimize free energy" putting researchers "in a position to consider behavior and self-organization" on the same basis.[40]

On this account familiar by now, agents reflect the environments in terms of which they are situated, with the dynamics of the world outside reflected in the structures inside of the input-output system at the center of which is the brain. Friston's thesis is that the brain works to maximize evidence for the model of the world which it embodies by acting on that evidence and testing it(self) against the perceptual reality. In minimizing surprise, the agent maximizes model likelihood to the point where endpoints of action are fully determined. This is to raise the question of why any agent would ever leave the safety of a fully determined situation at the risk of being surprised in the transition and suffering undue allostatic load, risking complete disintegration, a question addressed in terms of the "dark room problem." Briefly, given a sufficiently complex environment, the agent ventures forth because increasing information increases control in the long run such that opportunities to explore and to exploit new information add to the value of a given situation.[41] So as to why an agent might take risks, even seek them, it does so to maintain system integrity, so that the system does not dissipate in the face of entropic forces, and seeking—even creating—situations which best deliver security in the face of uncertainty: "the whole point of the free-energy principle is to unify all adaptive autopoietic and self-organizing behavior under one simple imperative; *avoid surprises and you will last longer.*"[42]

Consider the free-energy principle in the context of consciousness and minimal self. In a recent review of the field, Limanowski and Blankenburg trace the "minimal self" and its characteristic sense of mineness and ownership that we found at the heart of h-consciousness in our first paper through the early phenomenology of the twentieth century and in the form of a "self-model." On this view, "the agent

is the current embodied model of the world."[43] And as with Merleau-Ponty's "body-schema,"[44] minimal selfhood and the feeling that comes with it arises as a whole, with prediction of incoming sensory input and its influence on all levels of the self-model at once. The sense of mineness is thus "always *implicit* in the flow of information within the hierarchical generative self-model"—echoing Friston—"experienced for actions and perceptions in the same way." Accordingly, self is "not a static representation" but "the result of an ongoing, dynamic process" with the mineness most characteristic of consciousness "situated in a spatiotemporal reference frame where prediction introduces the temporal component of "being already familiar" with the predicted input."[45] Surprise, thus, is its natural complement, indicating subjective failure rather than merely objectively bad information.

Similarly, O'Regan develops the view that feelings derive from sensorimotor interaction with the environment. So long as there is interaction, then there is something that it is like to be so interacting, with consciousness arising as an agent "with a self" has "conscious access to the ongoing sensorimotor interaction."[46] He distinguishes three levels of self in terms of which artificial agents may be evaluated. First, the agent "distinguishes itself from the outside world." Second, "self-knowledge" expresses "purposeful behavior, planning and even a degree of reasoning." And, the third level is "knowledge of self-knowledge"—i.e., Goertzel's "reflective consciousness"—heretofore a "human capability, though some primates and possibly dogs, dolphins and elephants may have it to some extent."[47] O'Regan is optimistic that all three levels can be instantiated in AI. The question remains, how?[48]

On O'Regan's analysis, self is maintained under social forces which stabilize it as a construct, existing as a convenient figment like money. On his account, without the presumed value of money, the financial economy would fail and similar would hold for society in general should the value of "I" be doubted. People traffic in selves, in identities, because without it social order would disintegrate, i.e. surprise would not be minimized:

> Like the cognitive aspect of the self, the sense of "I" is a kind of abstraction that we can envisage would emerge once an agent, biological or non-biological, has sufficient cognitive capacities and is immersed in a society where such a notion would be useful.[49]

This "I" becomes useful when it relates personal experiences with others similarly situated, trading in information about what is worth having information about through the generalization of the self. This is a long way from pattern approximation, and farther away from identifying neural correlates with consciousness and self.

O'Regan's "I" captures the ubiquity of the self-model, but it fails to deliver just how this self-model comes to be constructed. What is missing is access to the dynamics that drive the formation of the self-model from the subjective perspective. This is because the structure of consciousness appears as only emergent phenomena. The idea is that consciousness is not a stable construct (like an "I") but appears during periods of relative instability through the circular causality developed among subjective mind, body, and environment. This circular causality cannot be captured in neural activity patterns alone, especially where these patterns are disrupted, and it cannot be expressed in terms of integration, as it is in disintegration and reintegration that consciousness emerges. Moreover, it cannot be captured in objective descriptions of "mineness" and of ownership of agency, as it is only for the agent itself that these descriptions are ultimately significant. Finally, as we shall argue in the next section, this is why synthetic neurorobotic experiments are necessary to access the essential structure of consciousness, as they offer a privileged perspective on the development of internal dynamics that ultimately ground the generalization and self-report of experience.

Tani summarizes the findings of three neurorobotic experiments in terms of three levels of self roughly coincident with O'Regan's, namely "minimal self, social self, and self-referential self." The first accounts for appearances of minimal selves in a simple robot navigation experiment, the second for appearances of social selves in an imitation learning experiment between robots and human subjects, and the third for appearances of self-referential selves in a more complex skill learning experiment. The following review of these results will put us in a position to appreciate Tani's central thesis regarding the role of criticality in the emergence of self and consciousness, as well as the importance of formal consciousness as set out in the next section.

In Experiment 1, interaction between the bottom-up pathway of perception and the top-down pathway of its prediction was mediated by internal parameters which adapted by way of prediction error.[50] System dynamics proceeded through the incremental learning process by intermittently shifting between coherent phases with high predictability and incoherent phases with poor predictability. Recalling Heidegger's famous analysis of the hammer as its failure reveals its unconscious yet skilled employment, consciousness arises with the minimal self as the gap is generated between top-down anticipation and bottom-up perceived reality during incoherent periods.[51]

Interestingly in this experiment, system dynamics proceeded toward a critical state characterized by a relatively high potential for a large range of fluctuations, and so to a relatively high potential for incoherency, analogous to the self-organized criticality (SOC) of Bak et al.[52] Tani speculated that SOC emerges when circular causality develops among neural processes as body dynamics act on the environment and then the body receives the reaction from the environment, with system level-dynamics emerging from mutual interactions between multiple local processes and the external world. During the first experiment for example, changes in visual attention dynamics due to changes in environmental predictability caused drifts in the robot's maneuvers. These drifts resulted in misrecognition of upcoming landmarks, which led to modification of the dynamic memory stored in the RNN, affecting later environmental predictability. Dynamic interactions took place as chain reactions with certain delays among the

processes of recognition, prediction, perception, learning, and acting, reflecting the circular causality between the subjective mind and the objective world. This circular causality provides for self-organized criticality. By developing this structure, breakdown to an incoherent phase proceeds only intermittently rather than all-or-nothing (similarly, the IIT). At the same time, Tani's thesis is that the self appears as momentary in these periods. In this way, this experiment was uniquely able to access the structure of consciousness as it affords a privileged view on the transition through meta-stable and unstable states to relatively stable states in terms of which automatic, unconscious, though perhaps skilled agency is regained.

Experiment 2 extended this research, exploring characteristics of selves in a social context through an imitation game between a humanoid robot controlled by the RNNPB and human subjects. The RNNPB is characterized by its simultaneous processes of prediction and regression.[53] In the middle of the mutual imitation game, analogous to Experiment 1 above, the RNNPB spontaneously shifted between coherence and incoherence. Tani and colleagues surmised that such complexity may appear at a certain critical period in the course of developmental learning processes in human subjects, when an adequate balance between predictability and unpredictability is achieved. Contrary to the image of a pinball simply following the paths of natural (nonliving) systems, human subjects may perceive robots as autonomous selves when these robots participate in interactive dynamics with criticality, as they actively self-determine possible ends and then test themselves in embodied action toward or away from them, pushing at the boundaries of the known and unknown in ways that other machines do not.

Experiment 3 addressed the problem of self-referential selves, i.e., does the robot have a sense that things might have been otherwise? Here, the RNNPB model was extended with hierarchy and as a neurorobotic arm manipulated an object, the continuous sensorimotor flow was segmented into reusable behavior primitives by stepwise shifts in the PB vector due to prediction error. Then, the higher level RNN learned to predict the sequences of behavior primitives in terms of shifts in this vector. Tani and colleagues interpreted the development of these dynamics as the process of achieving self-reference, because the sensorimotor flow is objectified into reusable units which are then manipulated in the higher level. When the sensorimotor flow is recomposed of such segments, it becomes a series of consciously describable objects rather than merely transitions between system states, a dynamic that may begin to account for how self-referential selves are constituted, such as when one takes an objective view of one's self as one "life story" among others.

That said, such constructs arising in this hierarchical RNNPB research cannot fully account for structures of self-referential selves. They are constituted in a static way, along a one-directional bottom-up path. Incidentally, experimental results using the same model regarding online plan modulation demonstrate how genuinely self-referential selves may be constituted.[54] These suggest that the sequencing of primitives in the higher level can become susceptible to unexpected perturbations, such as when an object is suddenly moved. Such perturbations could initiate critical situations. Due to the online nature of behavior generation, if the top-down expectations of PB values conflict with those from bottom-up regression, the PB vector can become fragmented. Even during this fragmentation, the robot continues to generate behaviors,but in an abnormal manner due to the distortion of the vector. The regression of this sort of abnormal experience causes further modulation of the current PB vector in a recursive way. During this iteration within the causal loop, the entire system may face intrinsic criticality from which a diversity of behaviors originates. And ultimately, this supports the contention that genuine constructs of self-referential selves appear with criticality through conflictive interactions in the circular causality of the top-down subjective mind and the bottom-up perceptual reality.

In summary, the three types of selves articulated above differ from each other, but more importantly they also share a similar condition of self-organized criticality that emerges in dynamic interaction between bottom-up and top-down processes. This condition cannot be accounted for by merely monotonic processes of prediction error minimization or free-energy, because such processes simply converge into equilibrium states (again, the dark room problem). Consciousness, and with it autonomy and the self cannot be explained in terms of convergent dynamics, but by ongoing open dynamics characterized by circular causality involving top-down prediction and bottom-up error regression, body dynamics acting on the environment and the reaction dynamics from the environment. Finally, in distinction from other research programs, Tani and colleagues' synthetic neurorobotics experiments are specifically designed to articulate these dynamics in a way that amounts to formal consciousness, as set out in the following section.

Recently, Tani examined free will arising from this open structure of consciousness by extending an MTRNN model to a scenario involving incremental interactive tutoring.[55] When taught a set of movement sequences, the robot generated various images as well as actions by spontaneously combining these sequences.[56] As the robot generated such actions, Tani occasionally interacted with the robot in order to modify its on-going movement by grasping its hands. During these interactions, the robot would spontaneously initiate an unexpected movement which Tani identified with an expression of free will. When Tani corrected the hand movement, the robot would respond by moving in yet a different way. Because the reaction forces generated between the robot's hands and Tani's hands were transformed into an error signal in the MTRNN, with its internal neural state modified through the resultant error regression, novel patterns were more likely to be generated when the robot was in conflict with the perceptual reality. The enactment of such novel intentions, experienced successively, induces further modification of the memory structure grounding further intention. Intentions for a variety of novel actions can thus be generated from such memory structures. And in this way, this experiment is able to isolate those dynamics grounding the emergence of free will in a synthetic neurorobotic agent.

In brief, the picture that emerges is that of a circular causality involving (1) spontaneous generation of intentions with various proactive actional images developed from the memory structure, (2) enactment of those actional images in reality, (3) conscious experience of the outcome of the interaction, (4) incremental learning of these new experiences and the resultant reconstruction in the memory structure.[57] Diverse images, actions and thoughts are potentially generated as the agent spontaneously shifts between conscious ("incoherent") and unconscious ("coherent") states with repeated confrontation and reconciliation between the subjective mind and the objective world. And summarily, free will as evidenced in the spontaneous generation of novel intention potentially arises as an open dynamic structure emerges through circular causality.

With this we see that self-reflective consciousness corresponding with O'Regan's third level may arise as an agent capable of revising intentions does so in order to meet a projected future situation according to self-determined plans to achieve it, in part by modulating its own agency by adopting predetermined or more reactive internal dynamics.[58] The ultimate question about the origins of an autonomous self becomes how subjective experience of continuous sensorimotor flow can be transformed into manipulable objects, memories and possibilities in terms of which self is both experienced and characterized. As the pure sensorimotor flow is segmented into identifiable objects, the flow in its original form becomes manipulable, and in its objectification becomes also generalized into an "I" stabilized through discourse with others similarly situated. Thus, Tani and colleagues' synthetic neurorobotics experiments have been able to isolate essential dynamics indicating self-organization through criticality to be the key mechanism driving the constitution of self-referential selves.

Our position is that self-referential selves emerge through self-organizing mechanisms involving the assembly and disassembly of sensorimotor schemata of repeated experiences, resulting in the construction of "self-models" or "body schemes" through internal dynamics. Most importantly, these arise only in *critical* conditions of sustaining conflictive and effortful interactions between the top-down subjective mind and the bottom-up sensorimotor reality at the level of agency. We cannot access consciousness in terms of a monotonic process of integration, error or free energy minimization, any more than through pattern matching and neural correlate tracking. For one thing, the ultimate aim of integrative dynamics is the "oneness with the world" which would characterize action without error within it. The result of this error free condition would, paradoxically by the present account, be consciousness of nothing at all. Rather, it is during purposeful conflict with the world that agent autonomy is exercised and self-consciousness arises, as it is against the silent standard of a perfect fit with project situations that an agent is held to account in inner reflection and correction of error. And moreover, it is due the structure of agency itself that the agent inherits from itself its own next situation at the end of each action, thereby cementing the "mineness" of h-consciousness that eludes being pinned down to any local neural correlate.

The preceding discussion shows that consciousness can be accessed by open dynamics where integration and breakdown are repeated during the exercise of agency in a changing world. Once again, pattern matching cannot afford such an insight, and in contrast with the IIT, consciousness appears when integrative dynamics break down. The essential structure of consciousness is the structure of autonomous agency simply put, a result that prepares us to appreciate the advance that Tani and colleagues' synthetic neurorobots represent in terms of formal consciousness in the following section.

3. INTRODUCTION TO FORMAL CONSCIOUSNESS

> What the soul nourishes by is of two types—just as what we steer by is both the hand and the rudder: the first both initiates motion and undergoes it, and the second simply undergoes it.
>
> – Aristotle[59]

Where the IIT holds that integration is essential to consciousness, with the integrative structure determining the phenomenal content of consciousness, and with "strong" integrated systems autonomous as they act and react from internally composed states and goals, Tani and colleagues' synthetic neurorobotic experiments show us how these goals are composed and why autonomy is necessary, in transitioning through critical periods toward relatively stable interactive states. This is a long way from where we began, at the door of Searle's Chinese room. And, it is in light of this advance that we wish to distinguish between "simulations" or "approximations" of consciousness and what we call "formal consciousness" instead, specifically in order to recognize Tani and colleagues' neurorobots as examples of the latter.

In Searle's Chinese room, there is an implicit interpretation of how AI works, what it does and how it does it, an interpretation that doesn't capture the essence of the neurorobots reviewed in this series of papers. His distinction between syntax and semantics is perhaps best understood to researchers in AI in terms of Steven Harnad's famous "symbol grounding problem,"[60] with much work in the direction of solving it since.[61] Let's reassess Searle's presumptions to better locate where we currently stand in the inquiry. Instead of merely matching incoming with outgoing symbols, the model agents reviewed in this series of papers anticipate input by forming appropriate output of its own prior experience, with the difference being used to refine that capacity going forward. This involves more than "input output behavior" as each input is transformed into something with strictly internal significance before output as something else with general significance. This is to say that the model develops its own private language, a phenomenon receiving recent popular attention in the context of AI[62] but which has been a long-standing point of interest in human beings.[63] This private language may be represented in terms of "patterns" and "shapes" but not directly, only after having been generalized and with the loss of the uniqueness that characterizes the deepest of human memories, so-called "flashbulb" memories for example. Still, a shared metric space mediated by common external objects grounds even these uniquely self-defining

memories in similar terms for those similarly situated, thus grounding generalization to common terms and facile communication of the significance of internal states so articulated.[64]

However, both private language and symbol grounding in a shared object environment neglect something fundamental to the phenomena of self, consciousness, and freewill, this being "how" this private language comes about as its limited grounds are exceeded and rediscovered through intermittent phases of incoherence. This dynamic has been emphasized in the preceding review of Tani and colleagues' neurorobotics. Their research formalizes the internal dynamics which not only facilitate translation from one grounded symbol to another, but that for example leave a human being hanging on a next word in anticipation. It is difficult to see how Searle's argument against first person ontology in an AI holds here. And, it is equally difficult to see how discovery of neural correlates of consciousness alone should reveal this fact. It may well be that conscious systems exhibit characteristic patterns in characteristic regions, but these may be duplicated without similar experience, "true zombies."

The models reviewed in this series of papers do not aim to duplicate neural correlates. Neither do they aim to simulate consciousness or to pass a Turing test. Rather, this research aims to isolate the essential structural dynamics in the normal operations of which certain phenomena arise. We refer to this aim as "formal" consciousness in distinction from others which aim at "reasonable approximations" evidenced in convincing behavior, for example. Specifically, we hold that three things are necessary for formal consciousness. First and foremost, there is critical reconciliation of intention with perceived reality as a system moves between relatively stable and unstable states, as discussed above.[65] This dynamic requires second that the system develop a private language which is then generalized into common terms through third a common grounding in a shared object environment. These three factors on the one hand account for unique subjectivity arising from otherwise common dynamic structures, while at the same time account for how this subjectivity and its uniqueness may be generalized in terms significant to other agents similarly situated. For human beings, this involves internalizing natural system energetics as a shared space of action, by way of which subjectivity can be "made sense of" by other human beings who are also grounded in this same object environment.[66] Note that this requirement is embodied in human beings as a product of evolution, and is captured by the FEP in current formal models which—in formal consciousness—stands in for the material component of biological consciousness, in this way opening the door to "making sense of" the experience of synthetic neurorobots in similar terms.

Formal consciousness represents the structural dynamics essential to consciousness, while simulated consciousness and reasonable approximations of behavior in Turing test capable so-called "general" AI need not. Here again, we may stress the point made in the first paper—this is a level of resolution that is inaccessible through study of biological consciousness—with the further caveat that not all synthetic models afford such insight, either. Only those designed to do so are able, instances of formal consciousness rather than something bent to a different end.

4. MOST- AND MYTH-CONSCIOUSNESS

> There is an originating and all-comprehending (principle) in my words, and an authoritative law for the things (which I enforce). It is because they do not know these, that men do not know me.
>
> – Tao te Ching, chapter 70, passage 2

Finally, we conclude with a short note on most- and myth-consciousness. Space forbids full exploration of this distinction, and in order to emphasize the role of criticality and incoherence in revealing the essential structure of consciousness, the following focuses on the promise for the current approach to formalize even the highest levels of human consciousness by way of dynamics common to the most basic.

There is precedent for distinction between levels of consciousness. For example, Gallagher distinguishes between pre-reflective and reflective consciousness in terms of minimal and "narrative" self.[67] Roughly in the first, an agent is aware of what it is undergoing, and in the second it recognizes such as episodic within the context of a "life story." The first is immediate though with an implicit sense of ownership, the "mineness" of h-consciousness as discussed in our first paper. The second is temporally extended, with episodes composed into stories that human beings tell about themselves and that come to define the self as fundamentally narrative in the strongest theories of narrative self. These can be mapped onto most- and myth-consciousness, with differences serving to clarify the point of the present distinction.

Most-consciousness corresponds with what IIT describes as the integration across differentiable system states, as in before and after the lights are turned on in a room. The felt difference between the two situations reveals the room. In so far as action proceeds according to expectation, there may be little in the sense of most-consciousness as in Tani's favorite example, making coffee without awareness of the process until after completion, when sitting with hot cup in hand reflecting on one's own apparent zombie-like activity and perhaps without capacity to self-report on the series of movements in between beyond prior generalization. This position is in concert with the phenomenological grounds of Gallagher's (2000) account of pre-reflective consciousness and its contrast with higher-order theories of consciousness on which consciousness arises with higher-order objectification of pre-reflective experience.[68] In terms of the neurorobots discussed in this series of papers, most-consciousness presents in the incoherence between predicted and perceived reality, for example when spilling the milk or dropping the spoon along the way, and includes the objectification of the movement that led to the mistake.

Most consciousness accounts for much, but it is not complete. To completely describe the feeling of what it is to be a self in a **maximal** sense, rather than in a minimal sense, we must describe what it feels like to generalize

the entire self, and not just one part and its possible actions. Gallagher attends to a similar phenomenon in the condition of "being a novelist" which on his assay involves "an enhanced ability for creating/entering into multiple realities and staying there longer and more consistently . . . without intersubjective support . . . short of dysfunction or delusion."[69] The novelist must create not only distinct narratives in the form of realistic life stories but also the coherent space of their interaction towards, ideally, some meaningful resolution. Myth-consciousness corresponds with this capacity, a sort of meta-narrative capacity that—on the present view—may be consequent on the experience of self-alienating criticality, an experience of a distance from one's own self-situation affording the experience of one's own entire self-situation as an object, and with this other self-situations as wholes similarly.

What may cause such deep criticality in a system that its subjective entirety may be taken as an object amongst others? We have introduced one possibility in the example of Aaron Applefield as discussed by Thomas Fuchs (2017) in the first paper. Trauma cementing memory "in the bones" in conflict with current perceptual reality may sustain the subject in the immersion in a perceptual reality that demands the "decoupling of conflict monitoring and executive control functions" which Gallagher proposes in novelists but that also confounds the "ability to re-connect and use executive control to come back to the default, everyday reality."[70] Such experience is also recognizable in the felt difference between one's present situation and that in which there is no self so situated at all, *angst*,[71] and which Victor Frankl (1985) understood contributes to the formation of purpose making the life of action as a whole meaningful. Myth-consciousness thus corresponds with what Gallagher discusses in terms of "delusion" and "dysfunction" understood as the normal function of a being aiming for coherence with an otherwise critically unstable situation, thereby discovering and indeed becoming the self-model of an underlying order that makes the transition to a relatively stable state, and the retention of personal integrity—even personal redemption—possible.

Our position is that the neurorobotic experiments reviewed in this series of papers formalize most-consciousness, and have not been designed for myth-consciousness, but are potentially myth-conscious. Currently, system criticality arises only at the moment of state instability and extends only to those local dimensions. However, myth-consciousness may be investigated through similar dynamics in an agent exposed to the necessary perceptual reality during sufficient personal development, e.g. human neoteny. Other approaches to artificial intelligence which focus on reproducing stable activity patterns for example cannot result in something like myth-consciousness, but it is a potential for synthetic neurorobotics as pursued by Tani and colleagues as an aspect of future research.

5. CONCLUSION

> If you have your *why*? for life, then you can get along with almost any *how*?
>
> – Nietzsche[72]

This series of papers made the case for formal consciousness in a family of neurorobots isolating dynamics essential to consciousness independent of neural correlates. It began with naturalistic nonreductionism and with consciousness in biological agents, resolving consciousness at the level of situated system open to the complex world, centering on the thesis that consciousness is a consequence of agential systems situated at the cusp of criticality, arising not in routine execution but in surprising failure to continue in perfect coherence with the world and thereby finding themselves out of place within it.

Tani and colleagues' synthetic neurorobots afford insight into the essence of consciousness where other systems cannot. They articulate the essence of free agency where other systems articulate something else to some other end. Finally, we may ask what it is that keeps us from understanding that consciousness inheres in such an artifact by design, even when confronted with products of consciousness at every turn? What is it that stops us from recognizing consciousness in an appropriately designed model intelligence, much as we recognize chairness in a chair, or computation in a computer? We answer that it is only our incapacity to recognize the origin of such phenomena in ourselves, in the reconciliation of the subjective with the objective world. As we reflexively aim for the restoration of stable coherency where otherwise there is only suffering, uncertainty, and the piercing awareness of it all, we retreat from conflict and away from the very object of our inquiry, away from consciousness itself. Without the courage to meet this struggle with a steady gaze, even with a machine articulating the truth of the matter, we fail to see it for what it is, formally conscious.

ACKNOWLEDGEMENTS

This paper, as the others before, has greatly benefited from Peter Boltuc's generous and insightful editing, as well as from his patient assistance alongside that of the staff responsible for producing this newsletter. The authors are deeply grateful to these people for their hard work, as well as for the guidance offered by anonymous reviewers in developing this series of papers. Thank you, all.

NOTES

1. Buszaki, Pyerache, and Kubie, "Emergence of Cognition from Action," 41.
2. Boltuc, "The Philosophical Issue in Machine Consciousness."
3. Searle, "Consciousness," 567.
4. Searle, "Minds, Brains, and Programs."
5. Pinar, et al., "Turing Test: 50 Years Later."
6. Exactly the sort of system that demands a Turing test.
7. Searle, "Consciousness," 576.
8. Ibid., 577.
9. Ibid., 568–69.
10. Searle points to Crick and Koch ("Consciousness and Neuroscience") and the notion that the neural correlates for all senses—the varied "building blocks of microconsciousnesses"—are integrated into "any conscious field" and that a science of consciousness starts by isolating these and working up (Searle, "Consciousness," 570; see also Crick and Koch, "Constraints on Cortical and Thalamic Projections: The No-Strong-Loops Hypothesis"). Compare with Driver and Spence: "subjective experience within one modality can be dramatically affected by stimulation within another" relative to "the presence of feedback pathways from convergence zones" such that "brain areas

traditionally considered as 'unimodal', may only be so in terms of their afferent projections" together producing a multimodally determined percept which nevertheless has the unimodal qualia associated with the activation of brain areas receiving afferent input from only one primary modality" ("Multisensory Perception: Beyond Modularity and Convergence," R734).

11. The notion of a unified field has deep roots in cognitive science. Consider, for example, Kurt Lewin's topological psychology (*Principles of Topological Psychology*), Heckhausen and Heckhausen's (*Motivation and Action*) account of motivational development, and Gallese's ("The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity") account of shared experience between similarly embodied agents via mirror neural activity in terms of a "shared manifold."
12. Searle "Consciousness," 574.
13. Llinas et al. ("The Neuronal Basis for Consciousness") and Tononi, Edelman, and Sporns ("Complexity and Coherency: Integrating Information in the Brain"). See, for example, Oizumi et al., "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," pp. 19–20, for a more recent explication of this position.
14. Tononi, "Consciousness as Integrated Information: A Provisional Manifesto," 239–40. There being no requirement for any exchange with the external world, either. Compare Howry, *The Predictive Mind*.
15. Tononi and Koch, "Consciousness: Here, There and Everywhere?," 13.
16. Tononi, "Consciousness as Integrated Information," 236.
17. Oizumi et al., "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," 19.
18. Tononi, "Consciousness as Integrated Information," 237.
19. Ibid.
20. Tononi and Koch, "Consciousness: Here, There and Everywhere?"
21. Goertzel, "Hyperset Models of Self, Will, and Reflective Consciousness," 51.
22. Uddin et al., "Functional Connectivity of Default Mode Network Components: Correlation, Anticorrelation, and Causality," for review. The first paper in this series emphasized the specific yet integral roles of various neural regions including the vmPFC in establishing a sense of a future into which an agent is more or less invested, and here we may emphasize also the role of this node in cognizing the activity of others in prospection as well as the internal simulation of experience, autobiographical remembering, and theory-of-mind reasoning (cf. Spreng & Grady, "Patterns of Brain Activity Supporting Autobiographical Memory, Prospection, and Theory of Mind, and Their Relationship to the Default Mode Network").
23. Heine et al., "Resting State Networks and Consciousness: Alterations of Multiple Resting State Network Connectivity in Physiological, Pharmacological, and Pathological Consciousness States."
24. Oizumi et al., "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," 21.
25. Ibid., 22.
26. Ibid., 21.
27. Tononi and Koch, "Consciousness: Here, There and Everywhere?," 13.
28. Tani, "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study," 421.
29. Ibid., 421.
30. Ibid.
31. As does O'Regan, "How to Build a Robot that Is Conscious and Feels"; see the next section of this paper for discussion.
32. Barandiaran, Di Paolo, and Rohde, "Defining Agency. Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action."
33. Tani, "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study."
34. Ibid., discussion on pages 422–24 and 440, respectively.
35. James, *The Principles of Psychology*.
36. "The sense of agency—a phenomenological consequence of enacting sensorimotor schemes," in "The Sense of Agency—A Phenomenological Consequence of Enacting Sensorimotor Schemes," 207.
37. Ozumi et al., "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," 21.
38. See Bubic et al., "Prediction, Cognition, and the Brain," for review.
39. Friston et al., "Free-Energy Minimization and the Dark-Room Problem," 6.
40. Ibid., 2. Friston endorses "dual aspect monism" on which internal states can be inferred from structures available to our inspection due to the processes that these structures undergo. Note the emphasis here on system-level dynamics.
41. Schwartenbeck et al., "Exploration, Novelty, Surprise, and Free Energy Minimization."
42. Friston et al., "Free-Energy Minimization and the Dark-Room Problem," 2. This is Friston's equivalent of the IIT's "strong integration" also raising the issue of how extensively an agent self-determines its interactive boundary.
43. Limanowski and Blankenburg, "Minimal Self-Models and the Free Energy Principle," 6.
44. Merleau-Ponty, *Phenomenology of Perception*.
45. Limanowski and Blankenburg, "Minimal Self-Models and the Free Energy Principle," 6.
46. O'Regan, "How to Build a Robot that Is Conscious and Feels," 133.
47. Ibid., 121.
48. Compare Goertzel ("Hyperset Models of Self, Will, and Reflective Consciousness"): self arises as a concept within the space of other concepts, "plausibly" alike what may be happening in human beings as patterns refer to themselves. However, limited to patterns alone, this raises the potential for an infinite regress, one that Goertzel recognizes. To avoid this problem, we must look further than patterns and at how and why they emerge i.e. through criticality such that there is no potential for such limitless reflection.
49. O'Regan, "How to Build a Robot that Is Conscious and Feels," 123.
50. Tani, "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study"; originally appearing in Tani, "An Interpretation of the 'Self' from the Dynamical Systems Perspective: A Constructivist Approach").
51. Heidegger, *Being and Time: A Translation of Sein und Zeit*.
52. Bak et al., "Self-Organized Criticality: An Explanation of the $1/f$ Noise."
53. Part 1 of Tani, "Autonomy of 'Self' at Criticality."
54. Ibid., originally appearing in Tani, "Learning to Generate Articulated Behavior through the Bottom-Up and the Top-Down Interaction Process."
55. Tani, *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*, discussion in Section 10.2.
56. As reviewed in Tani, *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*, Section 10.1.
57. Note the similarity of this cycle with that independently developed in White, *Conscience: Toward the Mechanism of Morality*; White, "Understanding and Augmenting Human Morality: An Introduction to the ACTWith Model of Conscience"; White, "An Information Processing Model of Psychopathy"; White, "Manufacturing Morality: A General Theory of Moral Agency

Grounding Computational Implementations: The ACTWith Model"; and White, "Models of Moral Cognition."

58. Murata et al., "Learning to Perceive the World as Probabilistic or Deterministic via Interaction with Others: A Neuro-Robotics Experiment."
59. *De Anima*, Book 2, chapter 4, 416b20, in Irwin and Fine, *Aristotle: Selections*, 187.
60. Harnad, "The Symbol Grounding Problem"; see also Harnad, "Can a Machine Be Conscious? How?"
61. For example, Dennett, "Consciousness in Human and Robot Minds." See also Stuart, 2010, for review.
62. For example, https://arxiv.org/pdf/1611.04558v1.pdf in the context of natural language translation and, more recently, https://arxiv.org/pdf/1703.04908.pdf in the context of cooperative AI.
63. Kultgen, "Can There Be a Public Language?"
64. On the current view, the function of the brain is as an extension of embodied cognitive agency, itself a special instance of physical systems. There is a syntax to the universe that prefigures human evaluation and ultimately grounds human semantics. Symbols are grounded more deeply than in an agent's object level interactions with the world. They are grounded in the way that these objects and the world itself works. Human semantics derive from this natural syntax, as agents internalize external dynamics in order to become the self-model that most assuredly retains integrity in the face of dissipative forces. In the models reviewed in this series of papers, this material ground is articulated in the free energy principle. In human beings, it is a matter of material embodiment.
65. Tani, *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*, section 10.2.
66. It is interesting to note the use of physical language to describe internal states, for example, in Lee and Schnall, "The Influence of Social Power on Weight Perception."
67. Gallagher, "Philosophical Conceptions of the Self: Implications for Cognitive Science."
68. Gallagher, "Phenomenological Approaches to Consciousness," 693.
69. Gallagher, "Why We Are Not All Novelists," 141, discussion beginning page 139.
70. Ibid., 139.
71. Heidegger, *Being and Time: A Translation of Sein und Zeit*.
72. Nietzsche and Large, *Twilight of the Idols*, 6.

REFERENCES

Aristotle, T. H. Irwin, and G. Fine. *Aristotle: Selections*. Indianapolis: Hackett Publishing, 1995.

Bak, P., C. Tang, and K. Wiesenfeld. "Self-Organized Criticality: An Explanation of the 1/*f* Noise." *Physical Review Letters* 59, no. 4 (1987): 381–84.

Barandiaran, Xabier, Ezequiel DiPaolo, and Marieke Rohde. "Defining Agency. Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action." *Journal of Adaptive Behavior* 10 (2009): 1–13.

Boltuc, P. "The Philosophical Issue in Machine Consciousness." *International Journal of Machine Consciousness*, 1, no. 1 (2009): 155–76.

Bubic, A., C. D. Yves, and R. I. Schubotz. "Prediction, Cognition, and the Brain." *Frontiers in Human Neuroscience* 4 (2010): 1–15.

Buhrmann, T., and E. DiPaolo. "The Sense of Agency—A Phenomenological Consequence of Enacting Sensorimotor Schemes." *Phenomenology and the Cognitive Sciences* 16, no. 2 (2017): 207–36.

Buzsáki, G., A. Peyrache, and J. Kubie. "Emergence of Cognition from Action." *Cold Spring Harbor Symposia on Quantitative Biology* 79 (2014): 41–50.

Crick, F., and C. Koch. "Consciousness and Neuroscience." *Cerebral Cortex* 8, no. 2 (1998a): 97–107.

Crick, F., and C. Koch. "Constraints on Cortical and Thalamic Projections: The No-Strong-Loops Hypothesis." *Nature* 391, no. 15 (1998b): 245–50.

Dennett, D. "Consciousness in Human and Robot Minds." In *Cognition, Computation, and Consciousness*. Oxford University Press, 1997. Retrieved August 13, 2017, from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198524144.001.0001/acprof-9780198524144-chapter-2.

Driver, J., and C. Spence. "Multisensory Perception: Beyond Modularity and Convergence." *Current Biology*, 10, no. 20 (2000): R731–R735.

Frankl, V. E. *Man's Search for Meaning*. New York: Washington Square Press, 1985.

Friston, K., C. Thornton, and A. Clark. "Free-Energy Minimization and the Dark-Room Problem." *Frontiers in Psychology* 3 (2012): 1–7.

Fuchs, T. "Self across Time: The Diachronic Unity of Bodily Existence." *Phenomenology and the Cognitive Sciences* 16, no. 2 (2017): 291–315.

Gallagher, S. "Philosophical Conceptions of the Self: Implications for Cognitive Science." *Trends in Cognitive Sciences* 4, no. 1 (2000): 14–21.

Gallagher, S. "Phenomenological Approaches to Consciousness," in *The Blackwell Companion to Consciousness*, edited by S. Schneider and M. Velmans, 686–96. Malden, MA: Blackwell, 2008.

Gallagher, S. "Why We Are Not All Novelists." In *Investigations into the Phenomenology and the Ontology of the Work of Art: What Are Artworks and How Do We Experience Them?* edited by P. F. Bundgaard and F. Stjernfelt, 129–43. New York: Springer, 2015.

Gallese, V. "The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity." *Psychopathology* 36, no. 4 (2003): 171–80.

Goertzel, B. "Hyperset Models of Self, Will, and Reflective Consciousness." *International Journal of Machine Consciousness* 3, no. 1 (2011): 19–53.

Harnad, S. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42, no. 1 (1990): 335–46.

Harnad, S. "Can a Machine Be Conscious? How?" *Journal of Consciousness Studies* 10 (2003): 67–75.

Heckhausen, J., and H. Heckhausen. *Motivation and Action*. Cambridge: Cambridge University Press, 2008. doi:10.1017/CBO9780511499821.

Heidegger, M., and J. Stambaugh. *Being and Time: A Translation of Sein und Zeit*. Albany: State University of New York Press, 1996.

Heine, L., A. Soddu, F. Gomez, A. Vanhaudenhuyse, M. Thonnard, V. Charland-Verville, et al. "Resting State Networks and Consciousness: Alterations of Multiple Resting State Network Connectivity in Physiological, Pharmacological, and Pathological Consciousness States." *Frontiers in Psychology* 3, Article 295 (2012).

Hohwy, Jakob. *The Predictive Mind*. Paw Prints, 2015.

Irwin, Terence, and Gail Fine. *Aristotle: Selections*. Indianapolis, IN: Hackett Publishing Company, 1995.

James, W. *The Principles of Psychology*, Vol. 1. New York, NY: Henry Holt, 1918.

Kant, I., and M. J. Gregor. *Practical Philosophy*. Cambridge: Cambridge University Press, 1996.

Kultgen, J. H. "Can There Be a Public Language?" *The Southern Journal of Philosophy* 6, no. 1 (1968): 31–44.

Lee, E. H., and S. Schnall. "The Influence of Social Power on Weight Perception." *Journal of Experimental Psychology General* 143, no. 4 (2014): 1719–25.

Lewin, K. *Principles of Topological Psychology*. New York: McGraw-Hill, 1936.

Limanowski, J., and F. Blankenburg. "Minimal Self-Models and the Free Energy Principle." *Frontiers in Human Neuroscience* 7 (2013).

Llinas R., U. Ribary, D. Contreras, and C. Pedroarena. "The Neuronal Basis for Consciousness." *Phil. Trans. R. Soc. London Ser. B* 353 (1998): 1841–49.

Merleau-Ponty, M. *Phenomenology of Perception*. Translated by C. Smith. London: Routledge and Kegan Paul, 1962.

Murata, S., Y. Yamashita, H. Arie, T. Ogata, S. Sugano, and J. Tani. "Learning to Perceive the World as Probabilistic or Deterministic via Interaction with Others: A Neuro-Robotics Experiment." *IEEE Trans. on Neural Networks and Learning Systems*, 2015. doi:10.1109/TNNLS.2015.2492140.

Nietzsche, F. W., and D. Large. *Twilight of the Idols, Or, How to Philosophize With a Hammer (Oxford World's Classics)*. Oxford University Press, 1998.

Oizumi, Masfumi, Larissa Albantakis, and Giulio Tononi. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0." *PLoS Comput Biol* 10, no. 5 (2014): e1003588.

O'Regan, J. K. "How to Build a Robot that Is Conscious and Feels." *Minds and Machines* 22, no. 2 (2012): 117–36.

Pinar, S. A., I. Cicekli, and V. Akman. "Turing Test: 50 Years Later." *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science* 10, no. 4 (2000): 463–518.

Schwartenbeck, P., T. FitzGerald, R. J. Dolan, and K. Friston. "Exploration, Novelty, Surprise, and Free Energy Minimization." *Frontiers in Psychology* 4 (2013): 1–5.

Searle, J. R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417.

Searle, J. R. "Consciousness." *Annual Review of Neuroscience* 23, no. 1 (2000): 557–78.

Spreng, R. N., and C. L. Grady. "Patterns of Brain Activity Supporting Autobiographical Memory, Prospection, and Theory of Mind, and Their Relationship to the Default Mode Network." *Journal of Cognitive Neuroscience* Volume 22, no. 6 (June 2010): 1112–23.

Tani, J. "An Interpretation of the 'Self' from the Dynamical Systems Perspective: A Constructivist Approach." *Journal of Consciousness Studies* 5, no. 5-6 (1998): 516–42.

Tani, J., and S. Nolfi. "Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-Motor Systems." *Neural Networks* 12, no. 7 (1999): 1131–41.

Tani, J. "Learning to Generate Articulated Behavior through the Bottom-Up and the Top-Down Interaction Process." *Neural Networks* 16 (2003): 11–23.

Tani, J. "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study." *Journal of Consciousness Studies* 11, no. 9 (2004): 5–24.

Tani, J. "Autonomy of 'Self' at Criticality: The Perspective from Synthetic Neuro-Robotics." *Adaptive Behavior* 17, no. 5 (2009): 421–43.

Tani, J. *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. New York: Oxford University Press, 2016.

Tononi, G. "Consciousness as Integrated Information: A Provisional Manifesto." *Biological Bulletin* 215, no. 3 (2008): 216–42.

Tononi, G, G. Edelman, and O. Sporns. "Complexity and Coherency: Integrating Information in the Brain." *Trends Cogn. Sci.* 2, no. 12 (1998): 474–84.

Tononi, G., and C. Koch. "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370, no. 1668 (2015): 1–18.

Uddin, L. Q., K. A. M. Clare, B. B. Biswal, C. F. Xavier, and M. P. Milham. "Functional Connectivity of Default Mode Network Components: Correlation, Anticorrelation, and Causality." *Human Brain Mapping* 30, no. 2 (2009): 625–37.

White, J. B. *Conscience: Toward the Mechanism of Morality*. Columbia, MO: University of Missouri–Columbia, 2006.

White, J. "Understanding and Augmenting Human Morality: An Introduction to the ACTWith Model of Conscience." In *Model-Based Reasoning in Science and Technology. Studies in Computational Intelligence, Vol. 314*, edited by L. Magnani, W. Carnielli, and C. Pizzi, 607–21. Springer: Berlin/Heidelberg, 2010.

White, Jeffrey. "An Information Processing Model of Psychopathy." In *Moral Psychology*, edited by Angelo S. Fruili and Luisa D. Veneto, 1–34. Nova Publications, 2012.

White, Jeffrey. "Manufacturing Morality: A General Theory of Moral Agency Grounding Computational Implementations: The ACTWith Model." In *Computational Intelligence*, edited by Floares, 1–65. Nova Publications, 2013.

White, J. "Models of Moral Cognition." In *Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues*, edited by L. Magniani, 363–91. Springer: Berlin/Heidelberg, 2014.

INTERVIEW

Cognitive Engines Contemplating Themselves: A Conversation with S. L. Thaler

Stephen L. Thaler
IMAGINATION ENGINES INC., ST. LOUIS

Kristen Zbikowski
HIBBING COMMUNITY COLLEGE

BACKGROUND

For the past thirty years, Stephen Thaler's work has been in the development of artificial neural networks (ANN). A major focus of his work has been to find a way to develop creativity within computers in a way that was more organic than the human-coded algorithms and rule sets used with sequential processing systems.

Thaler works with both less complex ANNs and the more sophisticated "Creativity Machines" (CM). ANNs are typically "single shot" in that a pattern propagates from inputs to outputs somewhat like a spinal cord reflex. They crudely model perception. Made recurrent they may serve as associative memories. In contrast, CMs are composed of multiple ANNs, contemplatively banging around potential ideas until an appropriate one is found.

Creativity Machines function via a process involving the interaction between two different types of neural networks, *imagitrons* and *perceptrons*. The imagitrons consist of internally perturbed ANNs that harness disturbances to their neurons and connections to create variations on stored memory patterns, generating potential solutions to posed problems. Once detected by unperturbed ANNs, the perceptrons, these solutions are reinforced as memories that can later be elicited by exciting or "perturbing" the imagitron at moderate levels.

The result of this process is that the imagitrons within CMs generate a succession of ideas making them functionally contemplative rather than reflexive. A self-monitoring aspect then comes from perceptrons "watching" this succession and selecting the most appropriate of these ideas. There are many internal processes involved, including the selective reinforcement of those notions having novelty, utility, or value.

The level of perturbation-induced stress to the system affects the type of "recall" the system produces. The more intense these disturbances within the system, the greater the error in reconstructing its stored memories, leading to false memories or confabulations. Too much stress causes the ANNs to produce too great a variation on reality and an eventual cessation of turnover of such candidate ideas. However, Thaler could adjust the stress level within the system to generate confabulations that were sufficiently novel and plausible enough to qualify as viable ideas. Even better, he could let other neural nets determine the novelty,