Agents of Habit: Refining the Artificial Life Route to Artificial Intelligence

Susana Ramírez-Vizcaya[1,2], Tom Froese[1]

[1]Embodied Cognitive Science Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Okinawa 904-0495, Japan

[2]Philosophy of Science Graduate Program, National Autonomous University of Mexico, University City, Coyoacan, 04510, Mexico City, Mexico

susana.ramirez@oist.jp

Abstract

We provide conceptual clues for one promising Artificial Life (ALife) route to Artificial Intelligence (AI) based on the notion of habit. We draw from an enactive approach that considers habits as the building blocks for mental life and, consequently, as the foundation for a science of mind. By taking this standpoint, this approach departs from the conventional view of intelligence in AI, which is based on "higher-order" cognitive functions. The first part of the paper addresses the idea of taking habits as the foundation for modeling intelligent behavior. This requires us to consider the so-called "scaling up" problem and rethink the concept of intelligence that still pervades in mainstream cognitive science. In the second part, we present the enactive approach to habits, emphasizing their adaptive and complex nature, as well as their fundamental role in guiding behavior. Finally, we acknowledge some limitations in the current enactive models of habits: either they are disembodied and decoupled, but allow for a rich landscape of attractors, or they are embodied and coupled, but remain too minimal. We propose a bridge between existing models and point to the need to go beyond the individual to include a social domain. We conclude that to better model intelligent behavior, embodied and situated agents must be capable of developing an increasingly complex network of habits from which an intelligent self emerges.

Introduction

This paper argues for an Artificial Life (ALife) route to Artificial Intelligence (AI) that takes habits as the departure point for understanding and modeling intelligent behavior. In this regard, we follow Egbert and Barandiaran (2014) in their proposal that the notion of habit "holds the potential to become a blending category between the biological and the psychological" (p. 2). ALife, it has been argued, can contribute to AI by providing a new starting point: What might be needed is a view of intelligence that radically departs from the one that has been held by orthodox cognitive science; one that takes seriously the continuity between life and mind instead of assuming that intelligent behavior consists of the deployment of so-called "higher-order" cognitive abilities (Pfeifer and Scheier, 1999). Our proposal takes precisely this stand. We think that this is the right time to reconsider this long-standing goal of ALife because, after a flurry of widely publicized technical advances in machine learning, there is a growing suspicion among commentators and researchers (see e.g., Froese and Taguchi, 2019; Mitchell, 2019; Simonite, 2019) that the quest for a more human-like AI, e.g. one capable of common sense, has arrived at a plateau again.

AI has gone through two major *AI winters* since its inception, in which exacerbated enthusiasm and over-optimistic promises were soon followed by increasing disappointment for unfulfilled promises and severe cuts in funding for AI research. After a new resurgence and boom of AI during the last two decades, history might repeat itself, since without new conceptual breakthroughs, it seems likely that the next AI winter is just around the corner. Mikolov (2020) even argues that,

> [i]n some sense, we are still living through the AI winter: although the popularity of the term 'AI' has increased greatly over the last years, a vast majority of the AI projects aim to solve very narrow, isolated tasks, with very limited efforts to define projects aiming on developing Human-level AI or AGI [Artificial General Intelligence] (p. 50).

As a response to the second AI winter in the early 1990s, ALife proposed itself as a suitable route to make AI go forward. A number of researchers, including Rodney Brooks, Francisco Varela, Luc Steels, Tim Smithers, and Christopher Langton gathered together in 1991 to discuss the proper direction that this route should take (Steels and Brooks, 1995). However, apart from isolated successes in bringing the principles and methods of ALife to AI research (e.g., Brooks, 1991; Maes, 1993), both fields still remain quite apart from each other. Renewed interest in working out this path has been shown since the *2018 International Conference on Artificial Life*, but the concrete route is still far from clear.

In this paper, we provide some conceptual coordinates towards this end based on a body of work on autonomous systems developed within enactive cognitive science. Rather than offering concrete solutions, we aim at signaling a promising route for future research on AI. The core idea is taking habits as "the most basic building blocks for mental life" (Egbert and Barandiaran, 2014, p. 3) and, consequently, as a potential foundation for a science of mind. This line of research has as a its conceptual background a long tradition in philosophy, psychology, and sociology that includes authors such as Hegel, Ravaisson, Husserl, Heidegger, Merleau-

Ponty, Dewey, Piaget, Bordieu, and Gibson, who have seen habits as an essential condition for making sense of and acting in the world.

In the first section, we sketch the proposal of taking habits as the foundation for modelling intelligent behavior, which is still an underdeveloped idea within ALife research. This idea means a radical departure from representationalist and neurocentric approaches to cognition, which conceive intelligence as depending on the capacity for building an internal model of the environment and making inferences and plans based on the information contained in that model (e.g., Lake et al., 2017). Instead, the approach presented here sees intelligence as the deployment of an embodied and situated know-how. This implies a continuity between what is commonly labelled "lower" and "higher" cognition, thus requiring us to address the so-called "scaling-up" problem.

The second section presents the enactive view of habits, which departs from the one that prevails in psychology and neuroscience, according to which habits are automatic and rigid responses to particular context cues. For the enactive approach, a habit is "*a self-sustaining pattern of behavior* that is formed when the stability of a particular mode of sensorimotor engagement is dynamically coupled with the stability of the mechanisms generating it". This means that a habit is both the result and the condition for its own enactment, since "the stability or recurrence of the behavior that the habit involves (smoking, reading, jogging) both depends-on and reinforces the mechanisms that give rise to it" (Barandiaran, 2008, p. 281). As we will see, this circularity makes habits stable, but not necessarily rigid, since they undergo a continuous process of equilibration that makes them adaptive and flexible. According to the enactive approach, habits self-organize into complex bundles that are frequently acted together in particular circumstances, giving rise to a global systemic identity or *self* (Barandiaran, 2008; Di Paolo et al., 2017; Ramírez-Vizcaya and Froese, 2019).

The last section presents an overview of current enactive models of habits and provides some critical comments that we hope can contribute to find a more suitable habit-based ALife route to AI. We argue that scaling up requires building embodied and situated agents able to generate and sustain complex networks of habits that eventually give rise to an intelligent self. Current enactive models either are embodied and embedded, but allow for oversimplistic tasks and behaviors, or are bigger yet remain disembodied and decoupled from the environment. We point to the need to find a bridge that integrates what we see as two different approaches to modeling habits: (1) agent-based models that use simulations of minimal mobile agents to investigate the dynamics of habits, and (2) self-optimizing models that use habituation to make complex networks find efficient solutions that satisfy several constraints.

From Habits to Intelligent Behavior

The construct of intelligence has been defined in a variety of ways throughout history and across disciplines, and no consensus has been achieved so far on how to define intelligence in general, and human intelligence in particular. Given AI's original goal of "building intelligent systems and to understand [sic] human intelligence" (Brooks, 1995, p. 57), this lack of agreement has also impacted AI research, making it difficult to achieve a widely accepted definition of its subject matter and scope (Wang, 2019). This is true even if the search for AI broadens to include non-human forms of intelligence.

Even if there is no agreement on the definition of intelligence, the view that has prevailed in AI in one form or other is that intelligence is related to abstract, in-the-head, "higher-order" cognitive capacities (e.g., reasoning, planning, expert knowledge, problem-solving, decision-making) that depend on inferential reasoning. If we take such a view, it is hard to see how a route from habits to intelligence can be established without having to add any extra load of mental representations and complicated mental gymnastics to the (supposedly) "light" hand luggage of habits. In this section, we argue that the proposed route from habits to intelligence is not only possible, but desirable if we aim at understanding intelligence and creating intelligent artificial systems. This will require us to rethink our conventional notions of both intelligence and habits.

Since its inception, ALife has argued for a reconceptualization of intelligence based not on inferences or information processing, but on the adaptive behavior of embodied and situated agents (see e.g., Pfeifer and Scheier, 1999; Steels and Brooks, 1995). This view of intelligence is arguably the one that AI requires for facing the current challenge of building autonomous systems able to operate and make decisions in open-ended, complex environments.

But is this view of intelligence enough for the AI goal of understanding and modeling human intelligence? Recent non-representational approaches to cognitive science have started to addressed the so-called *scaling-up problem*, i.e., the problem of explaining "higher-order" cognition in terms of adaptive sensorimotor interactions of an embodied autonomous agent with its changing environment (e.g., Di Paolo et al., 2019; Gallagher, 2017; Hutto, 2015; Kiverstein and Rietveld, 2018; Thompson, 2007). In so doing, they aim at blurring the sharp distinction between "higher" and "lower" cognition by scaling down the former. One way of attaining this is by showing that what are commonly known as "higher-order" cognitive operations are "elaborations and gradual complexifications that develop out of lower, non-representational forms of cognition" (Kiverstein and Rietveld, 2018, p. 148). We follow this literature —and a longstanding phenomenological tradition— in regarding intelligence as embodied and situated *know-how* that allows for adaptive behavior. As Gallagher (2017) suggests,

> [i]n contrast to this traditional, conceptualist, internalist conception of mind […], the alternative is to think of mental skills such as reflection, problem solving, decision making, and so on, as enactive, non-representational forms of embodied coping that emerge from a pre-predicative perceptual ordering of differentiations and similarities (p. 202).

We suggest that this "pre-predicative perceptual ordering of differentiations and similarities" results from the development and deployment of a repertoire of habits. Evolution can certainly shape organisms' morphology and nervous systems so that they be able to perform various cognitive tasks. The role of evolution in minimal adaptive behavior is explored, for

instance, in Beer's (2003) well-known model agents evolved for active categorial perception, in which artificial "evolution structures an agent's dynamics so that its behavior is appropriately sensitive or insensitive to the different classes of perturbation that it may encounter" (p. 237). However, in the case of humans, the ontogenetic learning involved in habit formation is crucial for an open-ended adaptive behavior. As we will discuss in the next section, habits allow agents to discern what is relevant for a concrete situation, i.e., what possibilities for action (i.e., affordances) they should be responsive to, as well as to see beyond the immediate situation to anticipate the outcomes of their actions. We take this ability to habituate in an adaptive manner to the complexities of the sociocultural world as an essential requirement for the development of common sense, and hence for human-like intelligence. A failure to meet this requirement is exemplified in the generalized version of the *frame problem* —i.e. the failure to detect what is relevant in a potentially infinite factual context— that AI still seems to face.

Although there are different formulations of the frame problem, it can be posed as that of designing an artificial agent able to generate reliable expectations (traditionally conceived in terms of a set of beliefs) about the relevant changes in the environment that its actions will produce, without having to take into account and explicitly discard all the non-relevant ones. A famous illustration of this problem comes from Dennett (1984). He imagines some mobile robots faced with the task of retrieving a power source from a room in which there is also an active bomb. A crucial part of their task is to realize that, as a side-effect, they will also take out a bomb that comes in the same wagon as the power source. Despite the many attempts to provide the imagined robots with the required theorems and algorithms, they keep failing to foresee the relevant side-effect and act in consequence.

Dennett hints a few times at the notion of habit while examining some implementations that have been devised for solving the problem (e.g., scripts that provide a system of habits on attention). However, these solutions are expressed in terms of how to internally represent (store) all the relevant information so that an artificial agent can use it when needed to make the right inferences and attend only to the relevant features, while assuming that some other features will remain as usual after it acts. Importantly, this information is understood in terms of propositional knowledge or "facts" about the world that the designer has to provide to the agent.

According to Wheeler (2005), this is precisely the stance that orthodox AI has taken towards the frame problem by assuming that "the agent, in order to generate online intelligence, needs to internally model contexts as sets of representational, knowledge-that states" (p. 181). This also applies to the related *commonsense knowledge problem*, which deals with the difficulty faced by AI designers when "attempting to represent, in anything other than an artificially restricted scenario, the kind of commonsense knowledge that humans display" (Wheeler, 2005, p. 177). This commonsense knowledge is usually conceived of as an internal model that integrates the facts about the world (again, knowledge-that) needed for the system to act appropriately in a situation. We agree with Wheeler in that

> any attempt to internally reconstruct the highly distributed and interconnected networks of involvements that constitute context-dependent significance, by building inner representations of those networks […], looks to be, at best, positively Herculean (p. 105).

Dennett (1984) recognizes the biologically implausibility of these kinds of solutions, since they require the designer to continually add new provisions every time the artificial agent has to deal with novel situations. Accordingly, Dennett doubts that such solutions offer "any plausible suggestions about how the backstage implementation of that conscious thinking is accomplished *in people*" (p. 146). However, he considers that the solution to the frame problem is to be found within an information-processing framework, in terms of a set of beliefs representing what the agent thinks to be true about the world. In a critique to Dennett's heterophenomenology, Dreyfus and Kelly (2007) challenge his assumption that, whenever a subject has a conscious experience, she has a belief. They appeal to the experience of affordances as a concrete example of the idea that "not all conscious experiences are beliefs" (p. 51). As we argue in the next section, we agree with these authors in that, in experiencing an affordance, "the agent *feels immediately drawn* to act a certain way […], he experiences the environment *calling for* a certain way of acting, and finds himself responding to the solicitation" (p. 52).

We propose that a way out of the frame problem and the commonsense knowledge problem is not to be found in the addition of more propositional knowledge or larger inference capacities, but on the practical knowledge (*know-how*) provided by a self-organized, adaptive network of habits. The general idea is that habits open up a space of action possibilities in which relevance is not an extra ingredient that a designer has to incorporate, or an artificial agent has to infer based on a representation of a previously neutral environment. Instead, relevance arises from a history of sensorimotor interactions that "settle into dispositions, skills, and knowledge" (Fuchs, 2018, p. 102). This know-how can be (and has been) understood (see e.g., Hutto, 2005) in terms of mental representations, internal models, or inferences (e.g., action-oriented representations and active inferences). However, as we will see in the next section, the way that it is understood here is in terms of dynamical sensorimotor processes of attunement in which the world forms a non-decomposable system with the agent's body (see, e.g., Di Paolo et al., 2017; Gallagher, 2017).

In the next section, we concentrate on the role that habits play in providing this know-how. For doing this, we draw mainly on recent proposals within enactive cognitive science that build on habits to account for the autonomy of cognitive agents, but we also integrate ideas from phenomenological research on body memory and skilled intentionality.

Habitual Identities and the Emergence of Relevance

One of the cornerstones of enactive cognitive science is the notion of autonomy, which was initially discussed at the biological level (Varela, 1979). For the enactive approach, an autonomous system is a self-organized network of processes (e.g., metabolic reactions in a cell or neuronal ensembles in the nervous system) that

> (i) recursively depend on each other for their generation and their realization as a network, (ii) constitute the system as a unity in whatever domain they exist, and (iii) determine a domain of possible interactions with the environment (Thompson, 2007, p. 44)

The most basic form of autonomous system is the living cell, which is defined as a unity in the biochemical domain through the recursive interdependency of a self-producing (*autopoietic*) network of metabolic processes. In this regard, it is said that autonomous systems have a self-constituted *identity* that distinguishes themselves from their surroundings. However, autonomous systems also need to exchange matter and energy with their surroundings and actively regulate such exchanges so as to produce the conditions for its continued existence. In the case of the cell, metabolic processes create a semipermeable membrane that acts as a boundary, both distinguishing the cell from its environment and regulating its interactions with it. In this way, the cell

> co-determines or co-specifies, out of an in-principle undifferentiated surrounding, the set of chemical components, physical parameter ranges (e.g. temperature and pressure), or types of perturbations that constitute its 'relevant' environment (Barandiaran, 2017, p. 411).

Relevance emerges out of the need of living systems to maintain their biological identity within certain limits of viability, i.e., out of the need to keep themselves alive. Accordingly, organisms do not inhabit a neutral world that has to be internally represented and infused with meaning. Rather, "a living being is immediately presented with a meaningful world", since "to live is to always be concerned with something, most fundamentally with the continuation of one's individual manner of living" (Froese, 2017, p. 34). In this sense, "[o]ne could say that the environment emerges from the world through the being or actualization of the organism" (Goldstein, 1963/1995, p. 26). However, given this theoretical perspective, how is it possible that relevance emerges for an artificial agent? Is biological autonomy a requirement for it? A practically plausible answer to this question lies in the notion of habit. As Di Paolo (2003) has suggested,

> [w]e may invest our robots not with *life*, but with the mechanisms for acquiring *a way of life*, that is, with habits. This may be enough for them to generate a natural intentionality, not based on metabolism, but on the conservation of 'one' way of life as opposed to 'another one' (pp. 31-32).

There has been a growing interest within enactive cognitive science in understanding how different levels of autonomy interpenetrate in the self-individuation processes of agents. An upshot of this is an incipient body of work that focuses on the study of habits as the basis for sensorimotor autonomy (e.g., Barandiaran, 2008; Di Paolo, 2005; Di Paolo et al., 2017; Egbert and Barandiaran, 2014; Ramírez-Vizcaya and Froese, 2019). The general idea is that just as metabolic processes self-organize in autonomous networks that constitute an identity in the biological domain (i.e., a living being), sensorimotor processes self-organize in autonomous networks that give rise to an identity in the sensorimotor domain (i.e., a sensorimotor agent). This is crucial for us here, since artificial agents might not need to self-produce at a metabolic level in order to be autonomous: they might generate a sensorimotor identity through the development of a self-sustaining organization of habits.

Habitual patterns of behavior result from an interdependent network of neural, bodily, and interactive processes whose functioning depends, in a circular fashion, on the frequent enactment of those same patterns of behaviors they contribute to generate (Egbert and Barandiaran, 2014). Habits are thus both the outcome and the cause of their constituting processes. According to Barandiaran (2017), "[t]his form of recursion makes it possible to understand a mild sense of identity for the habit, a locus of survival and a self-generating persistence" (p. 421). This recursion is what makes habits self-sustaining, since when the enabling conditions in the agent-environment system are in place, the whole network of processes maintains itself. But this is also what makes them precarious: they are always at risk of extinction if not continuously enacted. In analogy with the biological domain, "mental death occurs when continuous disruption of the sensorimotor coupling irreversibly destroys the capacity of the system to behave coherently" (Barandiaran, 2017, p. 424).

We said earlier that the preservation of an agent's organization at the biological level grounds the perception of relevance in her environment. In this regard, it grounds the *norms* that guide an agent's behavior: situations and doings are adequate or inadequate, good or bad, successful or unsuccessful for the agent if they contribute to or jeopardize the continuation of her biological identity. In a similar fashion, the new level of identity generation that appears in the sensorimotor domain grounds a new level of normativity: sensorimotor agents make sense of their environment and regulate their interactions with it based on the preservation not only of biological identities, but also of habitual ways of life. In the case that any difficulty (e.g., an injury, a disease or any transformation in the performance context or lifestyle) prevents the smooth execution of an organization of habitual sensorimotor patterns, agents will try to compensate for it, so that they can "*appropriately* enact the *right* sensorimotor coordinations on which a tangle of habits depends for its systemic equilibrium" (Barandiaran, 2017, p. 422).

Hence, particular situations and activities of sensorimotor agents "become meaningful not only in virtue of their contribution to biological survival, but also in virtue of their contribution to the stability and coherence of a sensorimotor repertoire" (Di Paolo et al., 2017, p. 39). For instance, a wave that an unexperienced swimmer would avoid is something that will offer plenty of opportunities for an experienced surfer to improve her surfing skills. Similarly, a series of paddle strokes will be appropriate if they provide the propulsion needed to catch a wave, but also if they fit the particular style of the surfer. In a different timescale, breathing exercises will be suitable for maintaining a surfer lifestyle, while the habit of smoking will be detrimental, since breath hold for long periods of time is an essential skill for surfing.

In this regard, the correctness of a particular action will depend on how well

> [w]e combine several sensorimotor engagements into a coherent whole. In other words, to perceive and act successfully, we must demonstrate certain sensitivities and certain mastery of circumstances. Our living bodies

must pre-reflectively understand how they move in the world and how the world changes in response (p. 76).

This mastery results from processes of habit formation and refinement that extend throughout the lifespan. It is through these processes that agents learn the sensorimotor regularities required for acting adaptively in a changing environment. It is also through these processes that agents incorporate the sociocultural norms that they should attend to in concrete situations. Importantly, such mastery does not necessarily involve the acquisition of propositional knowledge about sensorimotor contingency laws or sociocultural norms, though it might be required in some situations. What it certainly involves is the incorporation or sedimentation of a history of recurrent sensorimotor interactions "in an *embodied memory* of the objects and their affordances" (Fuchs, 2018, p. 115), which provides a pragmatic awareness of the action possibilities, experienced as a motivation to act (Butler and Gallagher, 2018). In the words of Merleau-Ponty (1945/2012), a "habit has been acquired when the body allows itself to be penetrated by a new signification, when it has assimilated a new meaningful core" (p. 148). Through the sedimentation of an interconnected set of habits, an experienced surfer will learn to recognize the precise moment and make the appropriate moves to catch unbroken waves. Likewise, after repeated experiences, a child will have learned to keep an adequate social distance in an elevator and a new resident to step into a public bus in Mexico City (if you have been there, you would probably know that it rarely fully stops).

Habits shape the agents' present and future engagements with the world and enable them to anticipate upcoming events and foresee the outcomes of their actions. For instance, in the act of reaching for a coffee mug while reading an article on the computer, my body anticipates its weight and form, making the correct movements even if the mug is lying on the margins of my perceptual field. For doing this, agents do not have to form internal representations of their environment and evaluate all the possible responses to it. Agents do not stand in a relation of representation with their environments, as two separate "systems that affect each other fundamentally via informational inputs and outputs" (Di Paolo et al., 2017, p. 35). Instead, they are in a relation of "ongoing dynamic coupling" (Fuchs, 2018, p. 102) in the course of which both agent and environment conform a system that is constantly reconfigured: "each interaction changes—even if only minimally— the structure and disposition of the organism that, in turn, perceives or reacts to its environment in a modified way" (p. 103).

A condition of possibility for the sedimentation of habits is the plasticity of the body —of its muscles, nerves, and cells. It is through "its capacity to be affected" (Carlisle, 2006, p. 26) that an agent is able to incorporate past experiences into its present acts. And this sedimentation, in turn, calls for the enactment of the sedimented habits in a circular fashion. According to Di Paolo et al. (2017),

> [t]his is analogous to the situation when people repeatedly walk across a lawn in a park, which leads, after some time, to the formation of paths where the grass is prevented from growing. This in turn encourages further walking along the paths, which continues to 'sediment' the path network. Similarly, a habit 'calls' for its exercise and its exercise in turn reinforces its durability (p. 144).

It might be objected that if the preservation of a habitual identity is what guides the behavior of an agent, it is not clear how this approach can account for novel behaviors and creativity. This is an important objection, since another

> key aspect, implicitly or explicitly present in many conceptions of intelligence, is generation of behavioral diversity while complying with the rules. […] An organism that always displays the same behavior is not intelligent (Pfeifer and Scheier, 1999, p. 32).

In the history of philosophy, habits often have been overlooked for being considered the antithesis of spontaneity and change. As Carlisle (2014) points out, authors such as Kant, Kierkegaard and Bergson have deemed habit as "an obstacle to reflection and a threat to freedom […] reducing spontaneity and vitality to mechanical routine" (p. 3). Even Ryle (1949/2009), who argued against an intellectualist view of cognition, regarded habit as the replica of an automatic, conditioned response devoid of intelligence. In psychology and neuroscience, habits are largely seen as the opposite of goal-directed behavior: habits may contribute to faster action control, but they lack the flexibility for adjusting to changing environmental circumstances (see, for instance, the literature on Reinforcement Learning). However, if we examine the dynamic organization of habits, we can see that they are not rigid and immutable (though they may become so, especially in pathological cases). Habits bring stability, but they are adaptive and open for change. Let us go briefly through this point before turning to the last section.

According to a dynamical systems interpretation of Piaget's theory of equilibration (Barandiaran, 2008; Di Paolo et al., 2017), individual habits and the whole sensorimotor organization undergo a continuous process of equilibration "by which a challenged agent-environment coupling may be adaptively steered back into its normal or into a new way of functioning" (Di Paolo et al., 2017, p. 84). Equilibration involves (1) the assimilation of small variations in the agent-environment system (e.g., perturbations, previously unencountered situations or objects with new properties), and (2) the accommodation of the supporting structures in the face of disruptions that impede the assimilation of new environmental dynamics, as Kohler's (1964) experiments with inverting or distorting prisms illustrate. This ongoing process of equilibration provides both flexibility and robustness (i.e., tolerance to noise and failure in its component processes) to an agent's behavior.

The capacity of habits to assimilate small variations makes it unlikely that each concrete enactment of a habit be identical, since the performance context and the internal dynamics of the agent will rarely be the same. Furthermore, as a result of accommodation processes, individual habits will adjust, transform, differentiate, or disappear, and new habits will be developed out of the previous ones. This process may lead to the re-equilibration of the whole network until it reaches a new state of (temporal) stability. This point is crucial, since according to the enactive approach, habits are "related to a plastic equilibrium that involves the totality of the organism, including other habits, the body, and the habitat they co-determine" (Barandiaran and Di Paolo, 2014, p. 5). Given that

habits ground a new level of normativity for the agent, a single, isolated habit might "take over the identity of the agent […], thus kidnapping, so to speak, the behavior generating mechanisms of the agent for its own perpetuation" (Barandiaran, 2008, p. 282). This would lead to the pathological repetition of one single sensorimotor pattern. However, in our daily life, a whole set of habits manifests simultaneously, both in the background and in the forefront, from walking and perceiving to reading and solving mathematical problems. As William James (1914) pointed out, living creatures are "bundles of habits". In the human case, some of them "are acquired as part of our sociocultural milieu, and some are idiosyncratic; all of them together reflect the history of each particular body." (Di Paolo et al., 2017, p. 81). This interdependence makes habits "more metastable (richer in potentialities) and adaptive than the traditional picture that associates habits with automatisms" (p. 147).

The interrelatedness of habits has also been acknowledged in the dynamical systems literature by Barrett (2014). Consistent with the enactive approach, this author considers habits as "the characteristic stability or stabilities of a system––its preferred states" (p. 1), which can be described through an attractor landscape that can change regularly in relation to various parameters. He suggests that, in learning, alterations to the attractor landscape in a particular behavior will alter "entire clusters of habits that are composed of shared components. That is, learning affects an entire 'habit space,' and not just individual habits in isolation" (p. 2). Though this author points to several ideas that are worth considering for further work in enactive research on habits, one that is relevant for this article —that is compatible with, but has not been explicitly worked out by the enactive approach—has to do with considering habits not only at the slower scales of months and years, but also at faster scales, with lower and faster levels showing a circular causality, so that "while habits of any given timescale are shaped by the 'deeper' habits of a slower timescale, they also can lead to changes at this deeper level" (p. 2). In this way, the idea of a whole network of habits broadens to include multiple levels in a temporal scale.

At this point of the journey, one might start wondering how this theoretical body of research on habits has been modeled under an enactivist approach. This is the topic that will occupy us in the following section.

Current Enactive Models of Habits

Given the relevance of habits for enactive cognitive science, it is surprising the scarcity of models that have been developed under this framework. In general terms, current enactive models of habits can be divided in two kinds. On the one hand, agent-based models use plastic controllers coupled to simulated mobile robots to investigate the formation and self-organization of habits, as well as their relation to basic metabolic autonomy. On the other hand, complex network models integrating a process of habituation have been used to find globally efficient solutions for satisfying general constraints. We will not go into the details of those models here. Instead, our purpose in this section is to provide a general description of these two kinds of modeling approaches, and to point to some of their limitations for modeling intelligent behavior.

The first agent-based models of habits were developed in the context of evolutionary robotics to study (1) the homeostatic adaptation of simulated mobile robots to radical sensorimotor disruptions (Di Paolo, 2000, 2003) and (2) the dynamics of behavioral preferences (Iizuka and Di Paolo, 2007). These models introduce a homeostatic neural controller that induces plastic local changes whenever neural activations go out of bounds, "until it finds a new internal dynamics which will make the system stable under the new conditions" (Di Paolo, 2000, p. 441). The general idea is that while neural dynamics are shaped by behavior, they also constrain behavior, making it more likely that it falls into certain attractors (i.e., habits). In the first case (Di Paolo, 2000, 2003), robots are evolved to perform phototaxis while maintaining their neural controllers within an homeostatic region. These robots manage to adapt to sensorimotor perturbations because, for them, "regaining internal stability also means performing phototaxis again" (Di Paolo, 2003, p. 16). In the second case (Iizuka and Di Paolo, 2007), two phototactic behaviors (i.e., approach light A or B) are associated through evolution to two regions of homeostatic stability. Once a robot selects a light, it tends to endure in that behavior, although some "spontaneous and externally-induced transitions" (p. 363) between preferences occur. This model thus serves as a minimal proof of concept for the idea that a habit is "sustained through time without necessarily being fully invariant, i.e., with time it may develop or it may be transformed into a different preference" (Di Paolo and Iizuka, 2008, p. 418).

These and subsequent models in evolutionary robotics using homeostatic adaptation (e.g., Aguilera et al., 2016; Iizuka et al., 2013) are important steps towards modeling sensorimotor autonomy "as a property of a system's organization" (Di Paolo and Iizuka, 2008, p. 409). However, they have some limitations as they remain rather neurocentric and with a tenuous link between neural homeostasis and adaptive behavior, since such association is established through artificial evolution. They also face the limitation that the homeostatic regions and the target behaviors associated with them are arbitrarily pre-defined and remain fixed over time. In this regard, Di Paolo and Iizuka (2008) acknowledge that "[i]t seems that having homeostatic regions that are somehow themselves constituted by a history of interactions would be a much better way of modelling autonomy" (p. 421).

Other agent-based enactive models of habits attempt to overcome these limitations by (1) focusing on the level of sensorimotor dynamics, while remaining agnostic on the details of biological realization, and (2) making the habitual patterns of behavior spontaneously emerge and sustain themselves through their recurrent enactment, instead of depending on "a pre-specified essential variable going outside of some pre-specified viability limits" (Egbert and Cañamero, 2014, p. 174). These models use a plastic habit-based controller, known as *itinerant deformable sensorimotor medium* (IDSM), coupled to a simulated mobile robot situated in an environment. The IDSM was developed by Egbert and Barandiaran (2014) to model the spontaneous "emergence and self-organization of habits" (p. 2). Variations of this controller have been used in other models to study, for instance, the relationship between habits and biological essential variables (e.g., Egbert and Cañamero, 2014) and the regulation of a

sensorimotor agent's interaction with the environment in response to self-generated norms (e.g., Egbert, 2018).

The basic idea is that the IDSM changes when the robot performs a particular behavior, making it more likely for the robot to be attracted to the sensorimotor (SM) trajectories that were more frequently traversed. In that way, "when a familiar SM-state is encountered, the IDSM produces behavior that is similar to the behavior that was performed when the agent was previously in a similar situation" (p. 4). This is a very minimal model that captures one of the core characteristic features of habits: their self-reinforcing character, in the sense that a behavior that was frequently enacted in the past is more likely to be re-enacted in the future —"perhaps in a slightly different form and provided that the environment continues to allow the SM-trajectory" (Egbert and Barandiaran, 2014, p. 8). This model does not only account for the stability of habits, but also for spontaneous changes when several habits have been formed. In one of the experiments, the IDSM is randomly initialized and different patterns of behavior emerge. While in some trails the robots stay endlessly in one habit, there are some cases where habits "naturally transition into another habit" (p. 10). We can see in this case that with the emergence of new habits, the system becomes metastable. However, as the authors recognize,

> [t]he question remains open as to whether a single habit is sufficient to speak of genuine autonomy and agency in the sensorimotor domain or a full self-regulating ecology of interrelated habits is required instead. Further variations and experiments with more complex environments, higher dimensional IDSMs or the addition of internal variables into the IDSM can be used to make progress in these or other directions (p. 12).

A few minimal models have recently explored how various conditions "bias the formation of habits so as to take different qualitative forms with different quantitative properties" (Zarco and Egbert, 2019, p. 585) and how "multistable systems with a greater variety of possible behaviours" (Woolford and Egbert, 2019, p. 8) emerge using very simple controllers. These are steps in the right direction. However, more work is needed to provide a proof of concept for the idea that "mental life emerges from a sensorimotor substrata that makes possible the development of an increasingly complex ecology of self-sustaining *sensorimotor* life-forms" (Egbert and Barandiaran, 2014, p. 13).

Philosophers such as Dewey, Merleau-Ponty and Deleuze have defended the idea that habits constitute the self, that "we not only *have* habits, but *are* habits" (Carlisle, 2006, p. 20). This means that "[s]elf-identity is maintained through time not by virtue of an unchanging entity, but through repeated action" (p. 23). The enactive approach to habits also pursues this idea through the notion of sensorimotor autonomy, proposing that a cognitive *self* emerges out of the self-organization of a network of habits (Barandiaran, 2008; Di Paolo, 2005, 2009, 2010; Di Paolo et al., 2017; Ramírez-Vizcaya and Froese, 2019). The idea is that habits self-organize in what Di Paolo (2009) calls "regional identities" that interact in complex ways. As we noted in previous work (Ramírez-Vizcaya and Froese, 2019), this idea can help us to understand how different sets of habits are deployed in particular spheres of an agent's life and how out of the "rich landscape of affordances" (Rietveld and Kiverstein, 2014) available to an agent, only a few of them stand out as relevant and solicit action depending on the regional identity at play.

Another limitation of current enactive models of habits that deserves a more in-depth analysis is their focus on individual agents, without taking into account the social dimension of habits. An exception to this is a recent model in evolutionary robotics that uses the homeostatic controller described above to understand the formation of social habits (Bedia et al., 2019). This is an important contribution to the enactive approach to habits, since, as multiple studies in developmental and social psychology have shown (and authors such as Dewey, Mauss and Bordieu have repeatedly stressed), there is a critical influence of the sociocultural environment in the development and deployment of even the most basic forms of habitual behavior. In this regard, Gallagher (2017) asserts that

> value distinctions between things in the environment that count as, and that we perceive as, salient or significant affordances (*versus* those that we don't) are laid out along affective, hedonic lines that are tied to other agents and what I see them do. Our perception of objects is shaped not simply by bodily pragmatic or enactive possibilities, but also by a certain intersubjective saliency that derives from the behavior and emotional attitude of others toward such objects (pp. 202-203).

These considerations lead us to the second kind of models: self-optimizing models that use habituation to make complex networks reach more optimal configurations that satisfy conflicting constraints. The first of these models was developed by Davies et al. (2011) in the context of game theory. Its purpose was to make "selfishly optimizing individuals" (p. 167) coordinate in such a way that they could optimize their (perceived) individual utilities while (almost) maximizing the global utility of the network. Since individual agents take into account only local information for acting, it is rather common that the constraints between individuals remain unsatisfied, resulting in a poor global performance. The key idea behind this model is that, "just like a neural network can optimize its collective dynamics via Hebbian learning […], a social network can optimize its collective dynamics via habitual learning" (Froese, 2018, p. 420). How habits develop within agents is not relevant for these models, since the focus resides on how habituation acts at the level of a social network. Through habituation, agents are made to slowly develop a preference for a coordination that may be suboptimal for them, but that is chosen because it is familiar. Habits are then repeatedly perturbed, allowing the system to settle into many different local attractors, which eventually "enlarges the basins of attraction for system configurations with high total utility" (Davies et al., 2011, p. 179). In this way, the system is able "to reach states of global utility higher than would have otherwise been possible" (p. 173).

This self-optimization model was latter used to study the hypothesized collective social organization of the ancient city of Teotihuacan, which was supposed to have faced the problem of reaching consensus between the multiple groups that shared the power (Froese et al., 2014). In this case, perturbations are proposed to correspond to "extreme community rituals involving temporary yet profound alterations of social relationships" (Froese, 2018, p. 421). A

similar model was also implemented at the neural level through an iterative process that uses Hebbian learning and random initializations to enlarge the basins of attraction of global optima (Woodward et al., 2015). In comparison to the agent-based models previously described, these models have the advantage of allowing for the interaction between multiple habitual behaviors. However, they are limited in that they are disembodied and decoupled from the environment. A proper route to AI would thus have to integrate these different approaches to modeling habits.

Conclusions

We have provided some conceptual clues that point to one promising ALife route to AI based on the notion of habit. By bringing together the enactive work on habits and some of the classical concerns in AI, we have also suggested a fruitful way to face the scaling-up problem in cognitive science: What might be needed for building an agent capable of common sense is a self-sustaining, autonomous network of regional identities from which a cognitive self can emerge. As Di Paolo and Iizuka (2008) have asserted, "[a] design process is now transformed into the design of the right conditions (appropriate material substrate and organization) for an autonomous identity to constitute itself" (p. 410). Current enactive models of habits offer some of the pieces needed to this end. Passing from a neural level of implementation to a sensorimotor one and exploring the development of social habits have been steps in the right direction. However, we need further work to increase the complexity of the modeled behaviors and integrate our social, symbolic, technological, and institutional context in this picture.

Acknowledgments. SR-V was supported by the Okinawa Institute of Science and Technology (OIST) and a scholarship granted by the Mexican National Council for Science and Technology (CONACYT).

References

Aguilera, M., Bedia, M. G., and Barandiaran, X. E. (2016). Extended neural metastability in an embodied model of sensorimotor coupling. *Frontiers in Human Neuroscience,* 10(76).

Barandiaran, X. E. (2008). *Mental life: a naturalized approach to the autonomy of cognitive agents.* Ph.D. thesis, UPV-EHU, University of the Basque Country.

Barandiaran, X. E. (2017). Autonomy and enactivism: towards a theory of sensorimotor autonomous agency. *Topoi,* 36: 409-430.

Barandiaran, X. E. and Di Paolo, E. A. (2014). A genealogical map of the concept of *habit*. *Frontiers in Human Neuroscience,* 8:522.

Barrett, N. F. (2014). A dynamic systems view of habits. *Frontiers in Human Neuroscience,* 8(682).

Bedia, M. G., Heras-Escribano, M., Cajal, D., Aguilera, M., and Barandiaran, X. E. (2019). Towards modelling social habits: an organismically inspired evolutionary robotics approach. In Fellermann, H., Bacardit, J., Goñi-Moreno, Á., and Füchslin, R. M., editors, *Proceedings of the Artificial Life Conference 2019*, pages 341-348. MIT Press, Cambridge, MA.

Beer, R. D. (2003). The dynamics of active categorial perception in an evolved model agent. *Adaptive Behavior,* 11(4): 209-243.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence,* 47(1-3): 139-159.

Brooks, R. (1995). Intelligence without reason. In Steels, L. and Brooks, R., editors, *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Butler, M. G. and Gallagher, S. (2018). Habits and the diachronic structure of the self. In Altobrando, A., Niikawa, T., and Stone, R., editors, *The Realizations of the Self*, pages 47-63. Palgrave Macmillan, Switzerland.

Carlisle, C. (2006). Creatures of habit: the problem and the practice of liberation. *Continental Philosophy Review,* 38: 19-39.

Carlisle, C. (2014). *On habit*. Routledge, London.

Davies, A. P., Watson, R. A., Mills, R., Buckley, C. L., and Noble, J. (2011). "If you can't be with the one you love, love the one you're with": how individual habituation of agent interactions improves global utility. *Artificial Life,* 17(3): 167-181.

Dennett, D. (1984). Cognitive wheels: the frame problem of AI. In Hookway, C., editor, *Minds, Machines and Evolution*, pages 129-151. Cambridge University Press, Cambridge, MA.

Di Paolo, E. A. (2000). Homeostatic adaptation to inversion of the visual field and other sensorimotor disruptions. In Meyer, J.-A., Berthoz, A., Floreano, D., Roitblat, H. L., and Wilson, S. W., editors, *From Animals to Animats VI: Proceedings of the 6th International Conference on Simulation of Adaptive Behavior*, pages 440-449. MIT Press, Cambridge, MA.

Di Paolo, E. A. (2003). Organismically-inspired robotics: homeostatic adaptation and teleology beyond the closed sensorimotor loop. In Murase, K. and Asakura, T., editors, *Dynamic Systems Approach for Embodiment and Sociality: From Ecological Psychology to Robotics*, pages 19-42. Advanced Knowledge International, Adelaide.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences,* 4: 429-452.

Di Paolo, E. A. (2009). Extended life. *Topoi,* 28: 9-21.

Di Paolo, E. A. (2010). Robotics inspired in the organism. *Intellectica,* 1-2(53-54): 129-162.

Di Paolo, E. A., Buhrmann, T., and Barandiaran, X. E. (2017). *Sensorimotor Life: An Enactive Proposal*. Oxford University Press, Oxford, UK.

Di Paolo, E. A., Cuffari, E. C., and De Jaegher, H. (2019). *Linguistic Bodies: The Continuity between Life and Language*. MIT Press, Cambridge, MA.

Di Paolo, E. A. and Iizuka, H. (2008). How (not) to model autonomous behaviour. *BioSystems,* 91: 409-423.

Dreyfus, H. and Kelly, S. D. (2007). Heterophenomenology: heavy-handed sleight-of-hand. *Phenomenology and the Cognitive Sciences,* 6(1): 45-55.

Egbert, M. (2018). Investigations of an Adaptive and Autonomous Sensorimotor Individual. In Ikegami, T., Virgo, N., Witkowski, O., Oka, M., Suzuki, R., and Iizuka, H., editors, *Proceedings of the Artificial Life Conference 2018*, pages 343-350. MIT Press, Cambridge, MA.

Egbert, M. and Barandiaran, X. E. (2014). Modeling habits as self-sustaining patterns of sensorimotor behavior. *Frontiers in Human Neuroscience,* 8:590.

Egbert, M. and Cañamero, L. (2014). Habit-based regulation of essential variables. In Sayam, H., Rieffel, J., Risi, S., Doursat, R., and Lipson, H., editors, *Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, pages 168-175. MIT Press, Cambridge, MA.

Froese, T. (2017). Life is precious because it is precarious: individuality, mortality, and the problem of meaning. In Dodig-Crnkovic, G. and Giovagnoli, R., editors, *Representation and Reality in Humans, Animal and Machines*, pages 33-50. Springer, Switzerland.

Froese, T. (2018). Ritual anti-structure as an alternate pathway to social complexity? The case of ancient Teotihuacan, Central Mexico. *Material Religion,* 14(3): 420-422.

Froese, T., Gershenson, C., and Manzanilla, L. R. (2014). Can government be self-organized? A mathematical model of the collective social organization of ancient Teotihuacan, Central Mexico. *PLoS ONE,* 9(10): 1-14.

Froese, T. and Taguchi, S. (2019). The problem of meaning in AI and robotics: still with us after all these years. *Philosophies,* 4(2): 1-14.

Fuchs, T. (2018). *Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind*. Oxford University Press, Oxford, UK.
Gallagher, S. (2017). *Enactivist Interventions: Rethinking the Mind*. Oxford University Press, Oxford, UK.
Goldstein, K. (1963/1995). *The Organism: A Holistic Approach to Biology Derived from Pathological Data in Man*. Zone Books, New York, NY.
Hutto, D. D. (2005). Knowing *what*? Radical versus conservative enactivism. *Phenomenology and the Cognitive Sciences,* 4: 389-405.
Hutto, D. D. (2015). Overly enactive imagination? Radically re-imagining imagining. *The Southern Journal of Philosophy,* 53: 68-89.
Iizuka, H., Ando, H., and Maeda, T. (2013). Extended homeostatic adaptation model with metabolic causation in plasticity mechanism —toward constructing a dynamic neural network model for mental imagery. *Adaptive Behavior,* 21(4): 263-273.
Iizuka, H. and Di Paolo, E. A. (2007). Toward spinozist robotics: exploring the minimal dynamics of behavioral preference. *Adaptive Behavior,* 15(4): 359-376.
James, W. (1914). *Habit*. Henry Holt and Company, New York, NY.
Kiverstein, J. and Rietveld, E. (2018). Reconceiving representation-hungry cognition: an ecological-enactive proposal. *Adaptive Behavior,* 26(4): 147-163.
Kohler, I. (1964). *The Formation and Transformation of the Perceptual World* (Vol. 3). International Universities Press, New York.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences,* 40: 1-72.
Maes, P. (1993). Modeling adaptive autonomous agents. *Artificial Life,* 1: 135-162.
Merleau-Ponty, M. (1945/2012). *Phenomenology of Perception*. Routledge, London.
Mikolov, T. (2020). Why is defining artificial intelligence important? *Journal of Artificial General Intelligence,* 11(2): 50-51.
Mitchell, M. (2019). How do you teach a car that a snowman won't walk across the road? *Aeon*. Retrieved from https://aeon.co/ideas/how-do-you-teach-a-car-that-a-snowman-wont-walk-across-the-road
Pfeifer, R. and Scheier, C. (1999). *Understanding Intelligence*. MIT Press, Cambridge, MA.
Ramírez-Vizcaya, S. and Froese, T. (2019). The enactive approach to habits: new concepts for the cognitive science of bad habits and addiction. *Frontiers in Psychology,* 10(301).
Rietveld, E. and Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology,* 26(4): 325-352.
Ryle, G. (1949/2009). *The Concept of Mind*. Routledge, London, UK.
Simonite, T. (2019). A sobering message about the future of AI's biggest party. *Wired*. Retrieved from https://www.wired.com/story/sobering-message-future-ai-party/
Steels, L. and Brooks, R. editors. (1995). *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*. Lawrence Erlbaum Associates, Hillsdale, NJ.
Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. The Belknap Press of Harvard University Press, Cambridge.
Varela, F. J. (1979). *Principles of Biological Autonomy*. Elsevier North Holland, New York.
Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence,* 10(2): 1-37.
Wheeler, M. (2005). *Reconstructing the Cognitive World: The Next Step*. MIT Press, Cambridge, MA.
Woodward, A., Froese, T., and Ikegami, T. (2015). Neural coordination can be enhanced by occasional interruption of normal firing patterns: a self-optimizing spiking neural network model. *Neural Networks,* 62: 39-46.
Woolford, F. and Egbert, M. (2019). Behavioural variety of a node-based sensorimotor-to-motor map. *Adaptive Behavior*(Special issue on Approaching Minimal Cognition): 1-16.
Zarco, M. and Egbert, M. D. (2019). Different forms of random motor activity scaffold the formation of different habits in a simulated robot. In Fellerman, H., Bacardit, J., Goñi-Moreno, Á., and Füchslin, R. M., editors, *Proceedings of the Artificial Life Conference 2019*, pages 582-589. MIT Press, Cambridge, MA.