FROM THE EDITOR, PETER BOLTUC

FROM THE CHAIR, DAN KOLAK

ARTICLES

SUSAN A. J. STUART
“Enkinaesthesia: Reciprocal Affective Felt Enfolding, a Further Challenge for Machine Consciousness”

GREGORY CHAITIN
“Metaphysics, Metamathematics, and Metabiology”

RICARDO SANZ
“Machines among Us: Minds and the Engineering of Control Systems”

THOMAS W. POLGER
“Toward a Distributed Computation Model of Extended Cognition”

COLIN ALLEN, JAIMIE MURDOCK, CAMERON BUCKNER, SCOTT WEINGART
“An API for Philosophy”
From the Editor

Piotr Boltuc  
University of Illinois at Springfield

On the first pages of his recent book Steven Hawkins' argues that philosophers, entangled in their narrow theoretical framework of methodology and philosophy of language, are no longer relevant in explaining the traditionally philosophical issues essential to understanding humans' role in the universe. In his note from the chair Dan Kolak seems to come forth with a response to those charges, while sharing some of Hawking's disappointment with our discipline. Dan puts philosophy and computers in the broader context of analytical philosophy and in a way, but only to a degree, as a nemesis of logicism predominant in its beginnings. Kolak views the interface between philosophy and computers as one important area where philosophers (to use a more radical phrasing than Kolak does) leave the ivory tower, alluded to by Hawkins, and their "philosophy schools" so as to reconnect with the world. The task requires a philosophy at the interface with the sciences "but not married to them." Referring to the articles in past and present issues of this Newsletter, and in particular those by Keith Miller and Greg Chaitin, Kolak says, "these little gems at the interface illustrate in a paradigmatic way how illuminating and deeply relevant a philosopher can be working at the borders between disciplines which, as has been pointed out time and again, is where all the interesting stuff happens."

We open the current issue with the article by Susan Stuart, a former active member of this Committee, on Enkinaestasia. This featured article highlights intersubjectivity as deeply co-affective and dependent on "the expressive, meaningful and cognitive bodily dynamic." Before reading this article I was always impressed by Susan's lectures but rather unable to understand her main philosophical message; the article at hand has changed this. It is followed by Gregory Chaitin's article in which he is not up for not a small task—he proposes the whole new field of metabolobiology, a general abstract (mathematical) theory of evolution "that studies the random evolution of artificial software (computer programs) rather than natural software (DNA)." Chaitin points out that "DNA is essentially a programming language that computes the organism and its functioning, hence the relevance of theory of computation for biology." The point is to model the space of all possible designs to see the role of evolutionary preassures; such modeling can occur only in the software space defined as "the space of all possible algorithms in a fixed programming language." Chaitin closes with a controversial claim. Based on an observation that human DNA is a bricolage of "bits and pieces of fish and amphibian design mixed in with that for a mammal," the author argues that we could do a much better job if we could design a human being out of scratch! Incidentally (talking about controversial claims!), both Chaitin and Stuart seem to belong to an exceedingly small group of serious authors who seem to question the standard interpretations, or applications, of the Church-Turing thesis.

The second block of articles comes from the invited session of this Committee at the Central Division APA meeting in Chicago in 2009. Ricardo Sanz, one of the top EU experts in robotics, discusses the issue of minds in the engineering of control systems. To put it somewhat crudely, Sanz argues that controlled systems require a mind since they need to understand what is going on. For instance, the reason why it is difficult to allow self-driven cars on the streets of our cities is that they do not understand what goes on, outside of the intricacies of the traffic. The fact that even not so smart human beings can drive reasonably well means that the solution is not to be adding detailed sub-routines preparing such artificial agents for more and more contingencies (this would be like adding epicycles to epicycles). The solution is for such systems to be provided with a different broader view when subroutines, like driving a car, figure out in a richer broad context of more human-like consciousness that incorporates "a world-model." In his article, Thomas Polger discusses A. Clark's extended mind hypothesis. He proposes "an alternative approach to extended cognition that does not depend on mere spatial and temporal extension." Instead, Polger argues that the right kind of constitutive coupling for an extended mind requires representations that are "distributed between the core system and the world, not just those that the system can 'loop' into the world to access."

We close the current issue with another note by Collin Allen and the team at Indiana University trying to build a unified ontology for philosophy. In the current article the authors discuss Application Programming Interfaces supposed to allow for broad argument mapping and present "the related ideas network" in the neighborhood of the Turing test.

In this Newsletter we say goodbye to Michael Byron and hello to Dan Kolak as the current chair of the APA Committee on Philosophy and Computers. Thank you Michael for all the support of this Newsletter, and of all the activities of our Committee.

Endnote

FROM THE CHAIR

Dan Kolak
William Paterson University

I’m grateful for this opportunity to offer a couple of reflections, a backward and forward looking one, about how philosophy in general and philosophy and computers in particular have an essential role to play in what traditionally has been the conundrum of broader philosophy of science—namely, the role of parent philosophy in relation to our science children.

When one looks back for instance on the clever sorts of schemes in those nifty early turns of twentieth-century philosophy of science by Carnap, Neurath, Morris et al., to create a platform of relevancy for philosophy at the interface of all the precocious advances in science and mathematics that at the time looked as if they were going to leave old funny dud philosophy in the dust, one cannot help but smile and shake one’s head, utterly at a loss. One wants to say, What were they thinking? But the kinder and more revealing response might be, What were they not seeing? Because we can answer that one the computer. Their use of sophisticated logical analysis of language in an attempt to make philosophy safe for the physical, the objective, the hard, while liberating philosophy from the mental, the subjective, the soft, as for instance was attempted in their apocryphal International Encyclopedia of Unified Science, only further alienated science from philosophy. There is some irony that while these fine logicians (including the good Russell, whose Principia tipped off Gödel, Von Neumann, Turing...) were using philosophy’s technical toys to play language games, scientists were using them to build bigger and better toys—like these little babies through which we’re all communicating at the moment, that have practically taken over all our lives, more so even than our kids (as father of four I apologize—with envy—to any childless philosophers who may resent the metaphor.)

But of course the serious, and philosophically more interesting point is that all our nasty, hotly good old philosophical conundrums not only are back but in full swing, and nowhere more evident and in play than at the interface between philosophy and computers, as amply evidenced by the richness and quality of articles, past and present, in this Newsletter. To me these little gems at the interface illustrate in a paradigmatic way how illuminating and deeply relevant a platform of relevancy for philosophy at the interface of all these technologies at the global level. What fourth-generation (4G) systems engineers have in store for us integrates virtually all our many technologies (physical layer technologies, network architectures, traffic management, middleware and context-aware solutions, QoS, etc.), into a single framework. Such connectivity among individuals (no longer limited primarily to those living in developed nations) is bringing human experience to a pivot point for the first time in its evolution, namely, the actualization—beyond the abstract metaphysical frameworks of philosophical speculation—of what for lack of a better word we might call “global consciousness.” One of the gravest problems facing us in the coming decades (aside from the ever growing technological gap between economically developed and underdeveloped peoples) is that we’ve seen time and again how such crucial junctures in which we manage to transcend in one way or another the borders that enclose and separate us (language, society, technology, etc.), each (r)evolution is fraught with new dangers, bringing unforeseen challenges that tend by and large to outpace our humanistic, i.e., our ethical, development. We thus implement astounding transformations of our external environment through engineering technologies—the control of nature, the advancement of political, legal and economic systems, etc.—while leaving the internal aspect of the human being—consciousness itself, subjectivity—virtually at a loss. As individuals, we human beings have yet to achieve the kind of subjective, internal transformation of which the philosophers of old and even of today sometimes speak in spiritual terms, as if such advancement is out of reach and unattainable except in the domain of pipe dreams and religion (a revoltingly redundant construction, I know). People tend thus to hold on to vast philosophical systems, curiously devoid of science, mathematics, and technology, as a last vestige of humanism against a technological world. This is a terrible mistake, in my view, one of the last vestiges of the old world order—the one in which those “recent philosophers of old”
worked, during those early turns of the twentieth, trying to salvage our discipline—which must evolve beyond its present phase if we are going ever to achieve anything like the relevancy of philosophy which many of us believe is necessary for our survival as an interconnected species.

Thus, all our technological and scientific advances notwithstanding, what philosophers in general but philosophers of science in particular need to enable in parallel with our next generation engineering systems is the development of philosophies beyond the next generation, which cannot be divorced from science and technology but nor can they be married to them. We need, akin to Independence Friendly logic, something like an Independence Friendly Philosophy of Science. This metaphilosophical task force cannot succeed in the abstract through “philosophy schools” churning out (pseudo) “technical” articles predicated on sprucing or propping up with newfangled words the old metaphysical systems of the past, the “steady as she goes” purveyors of insights intellectually detached from the evolving conceptual developments in science and mathematics driven by the technologically constructed systems within which we subsist as such, integrated across neurological and computational networks.

Beware, my fellow philosophers: while our technologies are already 4G, our philosophies are still 1G. As our scientific colleagues continue to create technological wizardry nothing short of magic, we have in many ways yet to leave Plato’s cave.

---

**ARTICLES**

**Enkinaesthesia: Reciprocal Affective Felt Enfolding, a Further Challenge for Machine Consciousness**

**Susan A. J. Stuart**
*University of Glasgow*

> “First, we interact and co-experience; later, as persons, we construe experience.”
>  
>  
> — Stephen Cowley (2008)

**Abstract**

In this short paper I will introduce an idea which, I will argue, presents a fundamental additional challenge to the machine consciousness community. The idea takes the questions surrounding phenomenology, qualia, and phenomenality one step further into the realm of intersubjectivity but with a twist, and the twist is this: that an agent’s intersubjective experience is deeply felt and necessarily co-affective; it is *enkinaesthetic*, and only through enkinaesthetic awareness can we establish the affective enfolding which enables first the perturbation, and then the balance and counter-balance, the attunement and co-ordination of whole-body interaction through reciprocal adaptation.

1. Introduction

The field of machine consciousness is as vibrant as ever, continuing to address crucial and thorny issues like machine phenomenology (Chrisley 2009), phenomenal consciousness (Boluc 2009; Gamez 2009), qualia and conscious machines (Haikonen 2009; Ramamurthy & Franklin 2009) and emotions (Haikonen 2003; Vallverdu & Casacuberta 2008; Takanishi 2009). But there is one issue which it has not yet encountered or considered and the challenge it presents is fundamental. I imagine that it will be a challenge that current theorists feel they are, in some way, already addressing or to which they think an answer will be forthcoming as a result of the satisfaction of the phenomenology, qualia, and emotions problems. My sense is that this would be to seriously misconstrue the essentially co-affective dialogical nature of conscious experience, and to underestimate the magnitude and nature of relational felt experience.

In one sense the point that I will be making is very simple: it is that agential bodies are co-affective sensory-kinaesthetic systems which spill out into the world and into the lives of others. It is essential that we understand this spilling out in relation to an agent’s conscious engagement with its world, but with our assumptions about bodies and boundaries this spilling out may at first seem counterintuitive. Our natural assumption is to see the boundary of the body as the limit of our experiential world, but it is precisely the breach of this boundary that provides us with the possibility of experience in the first place; the skin, overrun with an abundance of receptors—sixty kilometers of nerve fibers, fifteen kilometers of veins, with millions of sense receptors for pain, temperature, pressure, and touch—opens us up to the world and discloses it through our inescapable engagement with it (Hoffmeyer 2008), and the skin is supplemented by the plenitude of visual, proprioceptive, kinaesthetic, auditory, gustatory, and olfactory senses which open us up in their own way, are affected by change or motion within our world and which, with internal feedback, can bring about affective change within themselves.

Embodiment may be a nomological condition for agency (Dobyn & Stuart 2003) but it is the agent’s capacity to spill over into the bodily experience of others and vice versa, which establishes the community and reciprocity of felt co-engagement, and it is this felt co-engagement which is fleshed out in the expressive, meaningful, and cognitive bodily dynamics which are, in themselves, the necessary precursor to effective affective social, cultural, and linguistic communication in the human agent.

I will adopt the use of the neologism “enkinaesthesia” to refer to the reciprocal affective neuro-muscular dynamical flows and muscle tensions that are felt and enfolded between co-participating human (and, no doubt, animal) agents in dialogical relation with one another. Enkinaesthesia, like intersubjectivity and intercorporeality, relates to notions of affect, but in this case it is with the affect we have on the neuro-muscular dynamical flow and muscle tension of the other, including other animals, through our direct and our indirect touch. Direct touch includes the physical touch of a caress, a pat on the back, a hug, or the rebuff of the shrugged pulling away from contact. Indirect touch can be achieved through a look where one becomes the object of someone else’s subjective attention and experience, for example, in an unspoken admonishment, a papal blessing which can shrive you of your sins, a friend’s wave from a departing train, or in the way words and language, as biodynamical engines, can alter the way we feel.

2. The Feeling and Sensing Body

The feeling and sensing body has gained prominence in discussions of consciousness and experience in recent years, including the work of Damasio (1994, 1999 & 2003), Edelman (1992 & 2006), Ziemke (2003, 2007a & 2007b), Sheets-Johnstone (1999, 2000 & 2003) and, of course, a wealth of robotics work and, whilst I am generally sympathetic with these theories, they remain predominantly individual-centered and only
minimally-interactivist in character. Noë’s view (2004 & 2009) comes closest to my own, moving away, though not entirely, from the self-centered view, though he remains a little shy of the full commitment I want to make to the enkinaesthetic reciprocal affective neuro-muscular dynamical flow that is felt between agents in dialogical relation with one another. Noë writes:

The locus of consciousness is the dynamic life of the whole, environmentally plugged-in person or animal. Indeed, it is only when we take up this holistic perspective on the active life of the person or animal that we can begin to make sense of the brain’s contribution to conscious experience.

...Human experience is a dance that unfolds in the world and with others. You are not your brain. We are not locked up in a prison of our own ideas and sensations. The phenomenon of consciousness, like that of life itself, is a world-involving dynamic process.

(Noë 2009, p. xiii)

The moving, feeling, perceiving body is at the core of lived experience. But a non-relationally situated sensory-inkaesthetics with little consideration of the affectively laden interpersonal and interobjective world in which the agent finds itself will provide only a partial account of the experiential whole. Noë is right: the agent must be conceived from a holistic perspective, but the essential nature of the organism is not simply its kinaesthetic unfolding “in the world and with others”; the holistic perspective must embrace the agent not simply as a being in the world but as, and always as, a being with the world, folding into, enfolding with, and unfolding from those other agents and things with which it co-exists in utero to the point at which we depart this life. Ratcliffe (2008) speaks of this experiential entanglement as phenomenologically primitive:

World-experience is not distinct from how one’s body feels; the two are utterly inextricable. The experiential entanglement of body and world is more phenomenologically primitive than experience of either in isolation from the other. (Ratcliffe 2008, p. 1)

Lived experience is, first and foremost, enkinaesthetic.

3. Kinaesthesia and the Primacy of Movement

So let’s lay out the stall.

The cognitivist view of the mind, that presents the mind as symbolic, representational, and reducible to a set of physical states and processes that are fully explicable through scientific experiment and analysis, has been the predominant explanation for the mind in the second half of the twentieth century. At heart it is individual-centered and utilizes a substance-state ontology that treats temporality and spatiality as uniform, linear, and regular, consisting of discrete or punctuated events, points, objects, and places. On top of this it maintains the Enlightenment ideal of systematization—carving nature at its joints.

Enactivism, on the other hand, emphasizes the agent’s situation and embodiment in terms of its active, ongoing, processual, non-symbolic, non-representational based engagement in its world. It is essentially anti-dualistic, but unlike cognitivism’s inclination towards a monist materialism, the enactivist ontological commitments are rather more complicated. The agent is embodied and dynamically coupled to the world of other agents and things; thus, agent, world, and action are necessarily intricately interwoven, and the agent’s body, experience, action, and world together shape the way in which she deals with her everyday pragmatic concerns. Under this conception mind and world are inseparable, and it is embodied affective practice, rather than cognitive deliberation, that is the hallmark of the agent’s engagement with her world.

With only a slight modification enactivism embraces enkinaesthesia; the focal point moves from the agent and their individual agency to the necessity of our being co-agental in a co-dynamically continuous, affectively laden intersubjective and enkinaesthetic processual horizon of experience. “By a ‘way of finding oneself in the world,’” Radcliffe says, “I mean a sense of the reality of self and of world, which is inextricable from a changeable feeling of relatedness between body and world” (Radcliffe 2008, p. 2).

Thus it is that feeling bodies and things together in a dialogue of community and reciprocity with other feeling bodies and things play an integral role in full-bodied pre-linguistic sense-making relations.

Babies in the womb...send and receive messages without benefit of the words, syllables, and phrases that begin appearing in a year or two after birth. Their daily experiences of communication are punctuated by self-initiated and reactive movements which express needs, interests, and feelings. ...Based on the early development of the senses in the womb, a fetus remains in constant dialog with the surrounding environment. (Chamberlain 1995)

So, the genesis of this activity begins in utero and is necessarily co-agential, mother with prenate, occasionally mother with two or more prenates, and prenates with their bodies and the surrounding amniotic environment and beyond. “The maternal womb is an optimal, stimulating, interactive environment for human development. Activity never ceases and a fetus is never isolated,” and Chamberlain adds:

Between week six and ten, fetal bodies burst into motion, achieving graceful, stretching, and rotational movements of the head, arms and legs. Hand to head, hand to face, hand to mouth movements, mouth opening, closing, and swallowing are all present at 10 weeks (Tajani and Ianniruberto, 1990). By 14 weeks, the complete repertoire of fetal movements seen throughout gestation are already in evidence (deVries, Visser, and Prechtl, 1985). Movement is spontaneous, endogenous, and typically cycles between activity and rest. Breathing movements and jaw movements have begun. Hands are busy interacting with other parts of the body and with the umbilical cord.

From this early stage onward, movement is a primary activity, sometimes begun spontaneously, sometimes provoked by events. Spontaneous movement occurs earliest, probably expressing purely individual interests and needs. Evoked movement reflects sensitivity to the environment. For example, between 10 and 15 weeks g.a., when a mother laughs or coughs, her fetus moves within seconds. (Chamberlain 1997)

Our sensed and felt co-agency begins as soon as movement starts for this movement incorporates the sensations of touch, temperature, pain, hearing, balance, and orientation, chemosensors of smell and taste, mouthng, and sucking and licking, which are used to explore texture, hardness, and contours of objects, and, of course, the prenate’s own body and, in the case of twins, the other’s body too. Neither mouthing nor sucking and licking in this context are involved with eating and nutrition, rather they are, as are the others, affective dialogical means of exploration. The greatest advantage afforded the burgeoning agent is to feel as it moves and to move as it feels.
As Haikonen (2009) says about the post-natal agent, though he could just as well be speaking pre-natally:

Perception is an active inspection and exploration process that involves physical adjustments of the senses like eye and head motions and hand motions (touch). These motions result in accompanying kinesthetic information about body part positions and motions. (p. 230)

And I have, up until very recently, said much the same thing:

The sense of both an inner "egocentric" space (Brewer 1992) and an affective depiction—the sensation of being “out-there” (Aleksander and Dunmall 2003)—is formed through the rich interplay of the body’s sensory channels that receive information about the environment, its actuating system that enables manipulation of that environment, and its proprioceptive mechanisms which make it possible to sense the position, location, orientation and movement of the body and its parts. (Stuart 2009a)

But Haikonen is too reserved about the felt nature of the "kinesthetic information," probably wisely given the current lack of any real machine phenomenology, and both he and I omit to mention the enkinaesthetic phenomenology of the agent with their world, and it is this which makes the kinaesthetic information salient in the first place. It is only through enkinaesthetic awareness that the agent can establish the reciprocal affective enfolded required for the timely response and adaptation it needs to survive.

Thus, more recently I have written that:

We are deeply and naturally kinaesthetic and enkinaesthetic, aware of our bodily movement and our action in the world, but also able to affect others and be affected by them, moving and being moved (Bråten 2007) within a reciprocal affective neuro-muscular dynamical temporal flow. The way in which these felt somatosensory relations fold and unfold—by bringing forth our world through our kinaesthetic imagination and associated somatosensory expectations—together influences how we will shape and adapt our world, how we will then adapt to those changes, and so on. (Stuart 2009b, p. 179-80)

In agent-directed action, whether it is taking a step forward, reaching out tentatively with a hand, or gazing out over the landscape, we are continually, as part of our experiential horizon, asking tacit, non-propositionalized questions about our world and our being with and within it (Cotterill 1995 & 1998).

The feeling of being is, by its nature, a feeling of being with. Thus, the capacity for enkinaesthetic dialogue is an a priori nomological condition for agency and, through the creation of kinaesthetic memories, melodies and imagination (Stuart 2007 & 2009a), the generation of a felt anticipatory dynamics, making possible the effective engagement with object- and movement-dependent sensorimotor contingencies (Noë 2004). In our intersubjective openness we don’t just possess a transcendental intersubjectivity (Zahavi 1997), we possess a transcendental enkinaesthesia.

4. Enkinaesthesia

The enkinaesthetic dialogue is rarely, if ever, simply two, though with the influence that language has had on our thinking we do tend to characterize it in this way. We exist within an ongoing processual dialogue from our earliest moments in utero to the time in which we cease to feel, and at that point others don’t cease to feel, that is, be enkinaesthetically linked to us. This is part of a universal dialogue that consists of an innumerable web of relations of community and reciprocity of sensing and experiencing agents and things and their felt and, sometimes, explicitly intentional co-agency. It is this which co-constitutes conscious relations and the experientially recursive temporal dynamics of the non-symbolic, non-representationally based experiential horizon for all agents.

The organism does not develop in isolation from what happens around it; it is literally created (hence poiein) by nature, while at the same time modifying both nature and itself. In this respect, autopoiesis more accurately describes what in the phenomenological structure of Paarung is generally presented as an experiential circularity, because the former stresses that the autonomy of the living (being) is the very result of its contextual dependence. (Depraz 2008, p. 240)

Enkinaesthesia may emphasize the neuromuscular dynamics of the agent, the givenness (Henry 1963) of its experience, but it also emphasises the entwined, blended, and situated co-affective phenomenological structure of Paarung. Unlike the circularity of Paarung enkinaesthetic activity has a recursive dynamics, and it is these experientially recursive temporal dynamics that lead to the formation and maintenance of integral enkinaesthetic structures and melodies. Such deeply felt enkinaesthetic melodies emphasize the dialogical nature of the feeling of being as the feeling of being-with or being-among, and demonstrate the paucity of individuating notions that treat agents as singular.

If one wants to speak of a commitment to the alive consciousness of others here, one should speak not of a cognitive commitment but, rather, of a practical commitment. Like the baby in relation to her mother, we are involved with each other. It is our joint cohabitation that secures our living consciousness for each other. We live and work together. (Noë 2009, p. 33)

It is certainly our “cohabitation,” our being in affective relations of community and reciprocity, that secures our living consciousness for one another, and the pragmatics of the commitment, of the living and working together, are in a strong sense to do with survival. Describing it as a “practical commitment” emphasizes the bodily, kinaesthetic affective tonalities that underpin and make possible the proto-modal in relationships, what Gendlin calls the “implicit interactional bodily intricacy.”

There is an implicit interactional bodily intricacy that is first—and still with us now. It is not the body of perception that is elaborated by language, rather it is the body of interactional living in its environment. Language elaborates how the body implies its situation and its next behaviour. We sense our bodies not as elaborated perceptions but as the body sense of our situations, the interactional whole-body by which we orient and know what we are doing. (Gendlin 1992, p. 353)

Noë and Gendlin present compelling arguments, but their stated positions lack the reciprocal co-affectivity of these feeling states in the interactional dialogue. Such co-affectivity is characterized by being inherently intentional, which is to say that being-with and being-among is necessarily relational and comes already clothed in “aboutness.” The “knowing,” referred to by Gendlin, occurs through the enkinaesthetic affective enfolding which enables the balance and counter-balance, the
attunement and co-ordination of whole-body action through mutual reciprocal adaptation. It is this that Maturana refers to as “language,” communication which is fleshed out in the expressive, meaningful, and cognitive bodily dynamics.

At this stage we should begin to think about the implications of enkinaesthetic affective enfolding for the machine consciousness community.

Aleksander and Dunmall’s “depiction,” one axiom in a set of five that are together necessary for an agent A to be conscious of its sensorily-accessible world S, is said to present perceptual states that depict parts of S for A such that the agent has a context, an “out-there” which can be utilized in planning when and how the agent should act (Aleksander & Dunmall 2003). A natural agent’s transcendental enkinaesthesia makes a sense of both being with and being “out-there” a given in experience, for it is only through the touching and being touched, the experiential folding, enfolding, and unfolding that we sense our separateness but also our inseparability from our world, that is, our necessary co-agency within our world. This is the pre-paecious, pre-modal “knowing” to which Gendlin refers, the natural sense of our joint cohabitation according to Noë, and what Aleksander and Dunmall call the “having a private sense: of an ‘out-there’ world” (ibid., p. 8). But Aleksander and Dunmall admit that neither depiction alone, nor together with the other four axioms is sufficient for the artificial agent’s being conscious of or having that private sense of an “out-there” world.

According to Haikonen, “True conscious machines must have qualia, but the qualities of machine qualia need not be similar to the qualities of human qualia” (Haikonen 2009, p. 225). There seems little reason to disagree with this particular claim because it may be that proof one way or the other will always evade our grasp, but he does offer some guidelines for what could constitute some minimum requirements for machines with qualia:

• In order to facilitate qualia, do not make the system perceive via secondary symbols. Secondary symbols may be used in higher stages of cognitive processing.

• Use direct and transparent perception systems and integrate sensory and motor modalities seamlessly.

• Make the system inspect the world via explorative acts and let the products of the perception system and system reactions reflect the results of this inspection. (Haikonen 2009, p. 232)

These guidelines fit well with the pre- and post-natal non-propositionalized plenipotent exploration engaged in by a living agent but, as Haikonen himself says, “Unfortunately...this does not explain why and how exactly the feeling of...qualia can arise” (Haikonen 2009, p. 229). But this is precisely the rub. We do need to know how the feeling of qualia arise for the development of machine consciousness, for without qualia the nomological condition for deeply felt co-agency, the reciprocal affective neuro-muscular dynamical flows and muscle tensions that are felt and enfolded between co-participating agents in dialogical relation with one another, cannot be met. Without qualia there can be no phenomenologically primitive enkinaesthetic experiential entanglement, that is, no co-agency, and with no co-agency, there can be no conscious agency.

Conclusion
The concern in this article has been with the feeling and sensing body conceived in its crucial role within the non-individual-centered, enactivist dialogical nature of thought, mind, and agency. As Merleau-Ponty says, “(T)here is no inner man, man is in the world, and only in the world does he know himself” (Merleau-Ponty 1962, p. xii); we are always, without fail, in dialogue with our world: all action is interaction. We cannot act without our action being the result of processes that continue to move, shape, and direct us, and in our acting we move, shape, and direct the world: all action is reciprocal interaction. This dialogue is with objects and agents with which and with whom we are in a topologically complex web of dynamical, processual affective coordinative, orientational, intentional, and evaluative relations of community and reciprocity or, as Maturana states; we “operate in a domain of reciprocal coontogenic structural coupling through reciprocal structural perturbations” (Maturana 1988, §9.5). We are not simply “in” our world as individualized agents acting upon the other things in our world as though they are discrete entities, separate and separable from us; we are irreconcilably with and within our world, as much affected and effected by it as we effect and affect it.

We are endogenously intersubjective, folding enkinaesthetically into the being-in-time of the other. There are occasions when this enkinaesthetic engagement is not evident, for example, when someone’s behavior is pathologically unfeeling and unengaged, but these are the anomalies, the exceptions that prove the rule, and the subject of another paper altogether.

As Chrisley (2009) says, “An important goal of the field of machine consciousness is to make substantial contributions to the science of consciousness” (p. 53). In this paper I would hope to have done just that, not by saying that machine consciousness isn’t possible—I remain agnostic about that possibility—but by suggesting that to-date we have seriously misconstrued the essentially co-affective enkinaesthetic dialogical nature of conscious experience.

Afterthought
Let me clarify and respond to one possible objection to the notion of enkinaesthesia; enkinaesthesia is not heterophenomenology by another name. Heterophenomenology is simply the capacity to adopt an intentional stance with regard to others and to wait for some relevant empirical evidence by which their mentality can be verified. Enkinaesthesia has, I hope, been shown to be a great deal more than that.

Bibliography


Holland and Knight (2006), and Chella and Manzotti (2007) amongst a great many more.

For an interesting elaboration of how we can be affected by the look of another read Chapter 1 of Part 3 of *Being and Nothingness* by Sartre.


**Endnotes**

1. It’s possibly much more vibrant with the launch in June 2009 of the *International Journal of Machine Consciousness*, edited by Antonio Chella and published by World Scientific.


3. For an interesting elaboration of how we can be affected by the look of another read Chapter 1 of Part 3 of *Being and Nothingness* by Sartre.

4. That spoken words and language can act as "biodynamical engines" is Paul J. Thibault’s phrasing. Personal correspondence.

5. Direct touch may be straightforward to describe but experientially it is as vast and variable in effect and affect as indirect touch; the reason has to do with surfaces, boundaries, and borders, and what we perceive to be the limit of the bodily "self."

6. For a nice summary of embodied cognition work, though with a little too much emphasis on language for my own taste, see Borghi & Cimatti 2010.
7. This is just too numerous to mention but a good place to start would be Rodney Brooks’ work and, in particular, 1991a, 1991b & 1991c.
8. By “dialogical” I mean only the interactivity of agents and not textual, linguistic, or conversational activity.
9. From here on “sensory-kinaesthetics” will be encompassed in the term “kinaesthetic.”
10. For a commentary and discussion of enactive in utero development see Wood & Stuart 2009.
11. Clark provides the starkest example of an individual-centered cognitive approach in his Hypothesis of Organism-Centered Cognition (HOC): “Human cognitive processing (sometimes) literally extends into the environment surrounding the organism. But the organism (and within the organism, the brain/CNS) remains the core and currently the most active element. Cognition is organism centered even when it is not organism bound” (Clark 2008, p. 139).
13. It is at precisely this point that the “But you can’t tell that we’re not just brains-in-vats, that we only seem to be embodied” objections begin to pour in. In an earlier paper I and my co-author even defended the brain-in-a-vat hypothesis: “Now it is easy to construct a thought experiment in which a brain is disembodied in some nutrient bath and its afferent neural channels, sensory and proprioceptive, are given appropriate analogue stimuli, the process being controlled by a computer model of a 3-D world. More elaborately, impulses across the efferent actuating channel could be intercepted and fed into the computer model, feeding back into altered stimulation of the afferent channels to denote movement within, or change to, the world. To all intents and purposes, the unfortunate brain has a body—there is no way that the brain could tell it had not—but this body is not extended in physical space, only in virtual space” (Dobson & Stuart 2003, pp. 155-156), but we were mistaken to do so. As Meijsing (2006) demonstrates, we rather blithely assumed that the neural signals could do everything, not just sending the electrochemical messages along the nerve pathways but also somehow dealing with—a fact we ignored—the humoral signals, that is, the chemical messages that are sent by the bloodstream. Given the emphasis placed on the sensory system, the body, kinaesthesia, proprioception, and enkinaesthesia, in this present article, it would be nonsensical to overlook the crucial role of both types of message and the “body loop” which alters the “body landscape” (Damasio 1999, p. 54). Additionally, Cosmelli and Thompson (forthcoming 2011) urge us to think carefully about the biology of consciousness and what realizes subjective experience and, having done so themselves, they conclude that “Any vat capable of performing the necessary functions will have to be a surrogate body that both regulates and is regulated by the nervous system. In other words, the vat will have to exhibit a level of complexity at least as high as that of a living body with respect to bodily systems of life-regulation and sensorimotor coupling. Thus the entire system (vat plus brain) must satisfy these two basic requirements: (i) it must be energetically open and able to actively regulate the flow of matter and energy through it so as to control its own external boundary conditions (life-regulation); and (ii) it must be capable of actively regulating its own sensorimotor interactions with the outside world (sensorimotor agency). In short, the entire system must amount to a biologically autonomous, sensorimotor agent” (pp. 28-9).
14. It would certainly not be inconsistent at this stage to say that the enkinaesthetic action of the pre-natal infant, which establishes its “changeable feeling of relatedness” to its world, is the underpinning for later post-natal mirror activity, especially with regard to the third of the somatosensory neurons and the somatosensory proportion of the bimodal (visual as well) neurons occurring in the rostral part of the inferior parietal lobule. There certainly seems to be a strong case for saying that the affectivity of related feeling is ontogenetically prior to the affectivity of related seeing, that is, visual mirroring. See Gallese et al. 2002, and Rizzolatti & Craighero 2004.
15. Although “All of the sensory systems, except vision, need outside or exogenous stimulation as part of development in utero. The human visual system needs synchronous waves of retinal ganglion cell firing in utero but does not need light or vision” (Graven & Browne 2008, p. 171), there is good evidence of a pre-natal sensitivity to light: “The fetus can see at the end of the seventh month, and it reacts to changes in lighting and can follow a flashing light” (Kenner 2007, p. 228).
16. The concept of “environment” is used thickly to refer to the system’s world and its own variable internal states that are the subject of homeostatic functions.
17. We might understand self-givenness in terms of Husserl’s concept of “eidetic intuition”: the direct givenness which “refers to the acts in which ‘objects show up in person’” (Deprez et al. 2003, p. 45) and which primarily reveals itself as a perceptual and imaginative act concerned with disclosing an essence (ibid., p. 55). Self-givenness is concerned with the revelation of the tight experiential coupling between body and ownership of the experience.
18. Axiom 1 (Depiction):
A has perceptual states that depict parts of S.
Axiom 2 (Imagination):
A has internal imaginative states that recall parts of S or fabricate S-like sensations.
Axiom 3 (Attention):
A is capable of selecting which parts of S to depict or what to imagine.
Axiom 4 (Planning):
A has means of control over imaginative state sequences to plan actions.
Axiom 5 (Emotion):
A has additional affective states that evaluate planned actions and determine the ensuing action.
19. For further discussion of the elusive factor F see pages 15 and following in Aleksander & Dunnall 2003.
20. The same objection can be made to the ontogenesis of emotion and lack of phenomenology in Brown & Hussey’s article, “Emotional Cognitive Steps Towards Consciousness” (2009).
21. “Reciprocal” in this context does not imply equivalence of influence, feeling, or response; rather, to take a Kantian line on this, objects and agents, “so far as they coexist, stand in thoroughgoing community, that is, in mutual interaction” (Kant 1787/1929, A212, p. 233). The Kantian thesis of community and reciprocity of interaction carries with it the notion of “de-termination,” so that “Each substance... contain(s) in itself the causality of certain determinations in the other substance, and at the same time the effects of the causality of that other; that is, the substances must stand, immediately or mediately, in dynamical community” (ibid., A213/B260). Kant’s emphasis on “dynamical community” or commercium is exactly right, but his concern is with perception and time determination, whereas ours is with the enkinaesthetic affect which co-determines substances—agents and objects.
Metaphysics, Metamathematics, and Metabiology

Gregory Chaitin*
IBM Watson Research Center and University of Buenos Aires

Copyright by Gregory Chaitin.

Abstract
In this essay we present an information-theoretic perspective on epistemology using software models. We shall use the notion of algorithmic information to discuss what is a physical law, to determine the limits of the axiomatic method, and to analyze Darwin's theory of evolution.

Weyl, Leibniz, complexity, and the principle of sufficient reason
The best way to understand the deep concept of conceptual complexity and algorithmic information, which is our basic tool, is to see how it evolved, to know its long history. Let's start with Hermann Weyl and the great philosopher/mathematician G. W. Leibniz. That everything that is true is true for a reason is rationalist Leibniz's famous principle of sufficient reason. The bits of Ω seem to refute this fundamental principle and also the idea that everything can be proved starting from self-evident facts.

What is a scientific theory?
The starting point of algorithmic information theory, which is the subject of this essay, is this toy model of the scientific method:

theory/program/010 → Computer → experimental data/ output/110100101.

A scientific theory is a computer program for exactly producing the experimental data, and both theory and data are a finite sequence of bits, a bit string. Then we can define the complexity of a theory to be its size in bits, and we can compare the size in bits of a theory with the size in bits of the experimental data that it accounts for.

That the simplest theory is best, means that we should pick the smallest program that explains a given set of data. Furthermore, if the theory is the same size as the data, then it is useless, because there is always a theory that is the same size as the data that it explains. In other words, a theory must be a compression of the data, and the greater the compression, the better the theory. Explanations are compressions, comprehension is compression!

Furthermore, if a bit string has absolutely no structure, if it is completely random, then there will be no theory for it that is smaller than it is. Most bit strings of a given size are if it is completely random, then there will be no theory for it that is smaller than it is. Most bit strings of a given size are

Furthermore, Weyl attributes these ideas to Leibniz, to the 1686 Discours de métaphysique. What does Leibniz have to say about complexity in his Discours? The material on complexity is in Sections V and VI of the Discours.

In Section V, Leibniz explains why science is possible, why the world is comprehensible, lawful. It is, he says, because God has created the best possible, the most perfect world, in that the greatest possible diversity of phenomena are governed by the smallest possible set of ideas. God simultaneously maximizes the richness and diversity of the world and minimizes the complexity of the ideas, of the mathematical laws, that determine this world. That is why science is possible!

A modern restatement of this idea is that science is possible because the world seems very complex but is actually governed by a small set of laws having low conceptual complexity.

And in Section VI of the Discours, Leibniz touches on randomness. He points out that any finite set of points on a piece of graph paper always seems to follow a law, because there is always a mathematical equation passing through those very points. But there is a law only if the equation is simple, not if it is very complicated. This is the idea that impressed Weyl, and it becomes the definition of randomness in algorithmic information theory.²

Finding elegant programs
So the best theory for something is the smallest program that calculates it. How can we be sure that we have the best theory? Let's forget about theories and just call a program elegant if it is the smallest program that produces the output that it does. More precisely, a program is elegant if no smaller program written in the same language produces the same output.

So can we be sure that a program is elegant, that it is the best theory for its output? Amazingly enough, we can't: It turns out that any formal axiomatic theory A can prove that at most finitely many programs are elegant, in spite of the fact that there are infinitely many elegant programs. More precisely, it takes an N-bit theory A, one having N bits of axioms, having complexity N, to be able to prove that an individual N-bit program is elegant.

And we don't need to know much about the formal axiomatic theory A in order to be able to prove that it has this limitation.

What is a formal axiomatic theory?
All we need to know about the axiomatic theory A is the crucial requirement emphasized by David Hilbert that there should be a proof-checking algorithm, a mechanical procedure for deciding if a proof is correct or not. It follows that we can systematically run through all possible proofs, all possible strings of characters in the alphabet of the theory A, in size order, checking which ones are valid proofs, and thus discover all the theorems, all the provable assertions in the theory A.

That's all we need to know about a formal axiomatic theory A, that there is an algorithm for generating all the theorems of the theory. This is the software model of the axiomatic method studied in algorithmic information theory. If the software for producing all the theorems is N bits in size, then the complexity of our theory A is defined to be N bits, and we can limit A's power in terms of its complexity H(A) = N. Here's how:

Why can't you prove that a program is elegant?
Suppose that we have an N-bit theory A, that is, that H(A) = N, and that it is always possible to prove that individual elegant programs are in fact elegant, and that it is never possible to prove that inelegant programs are elegant. Consider the following paradoxical program P:

P runs through all possible proofs in the formal axiomatic theory A, searching for the first proof in A that an individual program Q is elegant for which it is also the case that the size of Q in bits is larger than
the size of P in bits. And what does P do when it finds Q? It runs Q and then P produces as its output the output of Q.

In other words, the output of P is the same as the output of the first provably elegant program Q that is larger than P. But this contradicts the definition of elegance! P is too small to be able to calculate the output of an elegant program Q that is larger than P. We seem to have arrived at a contradiction!

But do not worry; there is no contradiction. What we have actually proved is that P can never find Q. In other words, there is no proof in the formal axiomatic theory A that an individual program Q is elegant, not if Q is larger than P. And how large is P? Well, just a fixed number of bits c larger than N, the complexity H(A) of the formal axiomatic theory A. P consists of a small, fixed main program c bits in size, followed by a large subroutine H(A) bits in size for generating all the theorems of A.

The only thing tricky about this proof is that it requires P to be able to know its own size in bits. And how well we are able to do this depends on the details of the particular programming language that we are using for the proof. So to get a neat result and to be able to carry out this simple, elegant proof, we have to be sure to use an appropriate programming language. This is one of the key issues in algorithmic information theory, which programming language to use. 

Farewell to reason: The halting probability Ω

So there are infinitely many elegant programs, but there are only finitely many provably elegant programs in any formal axiomatic theory A. The proof of this is rather straightforward and short. Nevertheless, this is a fundamental information-theoretic incompleteness theorem that is rather different in style from the classical incompleteness results of Gödel, Turing, and others.

An even more important incompleteness result in algorithmic information theory has to do with the halting probability Ω, the numerical value of the probability that a program p whose successive bits are generated by independent tosses of a fair coin will eventually halt:

$$\Omega = \sum_{p \text{ halts}} 2^{-(\text{size in bits of } p)}$$

To be able to define this probability Ω, it is also very important how you chose your programming language. If you are not careful, this sum will diverge instead of being ≤1 like a well-behaved probability should.

Turing's fundamental result is that the halting problem is unsolvable. In algorithmic information theory the fundamental result is that the halting probability Ω is algorithmically irreducible or random. It follows that the bits of Ω cannot be compressed into a theory less complicated than they are. They are irreducibly complex. It takes N bits of axioms to be able to determine N bits of the numerical value

$$\Omega = .1101011...$$

of the halting probability. If your formal axiomatic theory A has $H(A) = N$, then you can determine the values and positions of at most N + c bits of Ω.

In other words, the bits of Ω are logically irreducible, they cannot be proved from anything simpler than they are. Essentially the only way to determine what are the bits of Ω is to add these bits to your theory A as new axioms. But you can prove anything by adding it as a new axiom. That's not using reasoning!

So the bits of Ω refute Leibniz’s principle of sufficient reason: they are true for no reason. More precisely, they are not true for any reason simpler than themselves. This is a place where mathematical truth has absolutely no structure, no pattern, for which there is no theory!

Adding new axioms: Quasi-empirical mathematics

So incompleteness follows immediately from fundamental information-theoretic limitations. What to do about incompleteness? Well, just add new axioms, increase the complexity $H(A)$ of your theory $A$! That is the only way to get around incompleteness.

In other words, do mathematics more like physics, add new axioms not because they are self-evident, but for pragmatic reasons, because they help mathematicians to organize their mathematical experience just like physical theories help physicists to organize their physical experience. After all, Maxwell’s equations and the Schrödinger equation are not at all self-evident, but they work! And this is just what mathematicians have done in theoretical computer science with the hypothesis that $P \neq NP$, in mathematical cryptography with the hypothesis that factoring is hard, and in abstract axiomatic set theory with the new axiom of projective determinacy.

Mathematics, biology, and metabiology

We’ve discussed physical and mathematical theories; now let’s turn to biology, the most exciting field of science at this time, but one where mathematics is not very helpful. Biology is very different from physics. There is no simple equation for your spouse. Biology is the domain of the complex. There are not many universal rules. There are always exceptions. Math is very important in theoretical physics, but there is no fundamental mathematical theoretical biology.

This is unacceptable. The honor of mathematics requires us to come up with a mathematical theory of evolution and either prove that Darwin was wrong or right! We want a general, abstract theory of evolution, not an immensely complicated theory of actual biological evolution. And we want proofs, not computer simulations! So we’ve got to keep our model very, very simple.

That’s why this proposed new field is metabiology, not biology.

What kind of math can we use to build such a theory? Well, it’s certainly not going to be differential equations. Don’t expect to find the secret of life in a differential equation; that’s the wrong kind of mathematics for a fundamental theory of biology.

In fact, a universal Turing machine has much more to do with biology than a differential equation does. A universal Turing machine is a very complicated new kind of object compared to what came previously, compared with the simple, elegant ideas in classical mathematics like analysis. And there are self-reproducing computer programs, which is an encouraging sign.

There are in fact three areas in our current mathematics that do have some fundamental connection with biology, that show promise for math to continue moving in a biological direction:

- Computation, Information, Complexity.
- DNA is essentially a programming language that computes the organism and its functioning; hence the relevance of the theory of computation for biology.

Furthermore, DNA contains biological information. Hence the relevance of information theory. There are in fact at least four different theories of information:
- Boltzmann statistical mechanics and Boltzmann entropy,
- Shannon communication theory and coding theory,
- algorithmic information theory (Solomonoff, Kolmogorov, Chaitin), which is the subject of this essay, and
• quantum information theory and qubits.

Of the four, AIT (algorithmic information theory) is closest in spirit to biology. AIT studies the size in bits of the smallest program to compute something. And the complexity of a living organism can be roughly (very roughly) measured by the number of bases in its DNA, in the biological computer program for calculating it.

Finally, let’s talk about complexity. Complexity is in fact the most distinguishing feature of biological as opposed to physical science and mathematics. There are many computational definitions of complexity, usually concerned with computation times, but again AIT, which concentrates on program size or conceptual complexity, is closest in spirit to biology.

Let’s emphasize what we are not interested in doing. We are certainly not trying to do systems biology: large, complex realistic simulations of biological systems. And we are not interested in anything that is at all like Fisher-Wright population genetics that uses differential equations to study the shift of gene frequencies in response to selective pressures.

We want to use a sufficiently rich mathematical space to model the space of all possible designs for biological organisms, to model biological creativity. And the only space that is sufficiently rich to do that is a software space, the space of all possible algorithms in a fixed programming language. Otherwise we have limited ourselves to a fixed set of possible genes as in population genetics, and it is hopeless to expect to model the major transitions in biological evolution such as from single-celled to multicellular organisms, which is a bit like taking a main program and making it into a subroutine that is called many times.

Recall the cover of Stephen Gould’s Wonderful Life on the Burgess shale and the Cambrian explosion? Around 250 primitive organisms with wildly differing body plans, looking very much like the combinatorial exploration of a software space. Note that there are no intermediate forms; small changes in software produce vast changes in output.

So to simplify matters and concentrate on the essentials, let’s throw away the organism and just keep the DNA. Here is our proposal:

Metabiology: a field parallel to biology that studies the random evolution of artificial software (computer programs) rather than natural software (DNA), and that is sufficiently simple to permit rigorous proofs or at least heuristic arguments as convincing as those that are employed in theoretical physics.

This analogy may seem a bit far-fetched. But recall that Darwin himself was inspired by the analogy between artificial selection by plant and animal breeders and natural section imposed by malthusan limitations.

Furthermore, there are many tantalizing analogies between DNA and large, old pieces of software. Remember bricolage, that Nature is a cobbler, a tinkerer? In fact, a human being is just a very large piece of software, one that is $3 \times 10^{10}$ bases = 6 × 10^9 bits = one gigabyte of software that has been patched and modified for more than a billion years: a tremendous mess, in fact, with bits and pieces of fish and amphibian design mixed in with that for a mammal.7 For example, at one point in gestation the human embryo has gills. As time goes by, large human software projects also turn into a tremendous mess with many old bits and pieces.

The key point is that you can’t start over, you’ve got to make do with what you have as best you can. If we could design a human being from scratch we could do a much better job. But we can’t start over. Evolution only makes small changes, incremental patches, to adapt the existing code to new environments.

So how do we model this? Well, the key ideas are:

Evolution of mutating software,

and:

Random walks in software space.

That’s the general idea. And here are the specifics of our current model, which is quite tentative.

We take an organism, a single organism, and perform random mutations on it until we get a fitter organism. That replaces the original organism, and then we continue as before. The result is a random walk in software space with increasing fitness, a hill-climbing algorithm in fact.8

Finally, a key element in our proposed model is the definition of fitness. For evolution to work, it is important to keep our organisms from stagnating. It is important to give them something challenging to do.

The simplest possible challenge to force our organisms to evolve is what is called the Busy Beaver problem, which is the problem of providing concise names for extremely large integers. Each of our organisms produces a single positive integer. The larger the integer, the fitter the organism.9

The Busy Beaver function of $N$, $BB(N)$, that is used in AIT is defined to be the largest positive integer that is produced by a program that is less than or equal to $N$ bits in size. $BB(N)$ grows faster than any computable function of $N$ and is closely related to Turing’s famous halting problem, because if $BB(N)$ were computable, the halting problem would be solvable.10

Doing well on the Busy Beaver problem can utilize an unlimited amount of mathematical creativity. For example, we can start with addition, then invent multiplication, then exponentiation, then hyper-exponentials, and use this to concisely name large integers:

$$N + N \to N \times N \to N^N \to N^{N^N} \to ...$$

There are many possible choices for such an evolving software model: You can vary the computer programming language and therefore the software space, you can change the mutation model, and eventually you could also change the fitness measure. For a particular choice of language and probability distribution of mutations, and keeping the current fitness function, it is possible to show that in time of the order of $2^N$ the fitness will grow as $BB(N)$, which grows faster than any computable function of $N$ and shows that genuine creativity is taking place, for mechanically changing the organism can only yield fitness that grows as a computable function.11

So with random mutations and just a single organism we actually do get evolution, unbounded evolution, which was precisely the goal of metabiology!

This theorem may seem encouraging, but it actually has a serious problem. The times involved are so large that our search process is essentially ergodic, which means that we are doing an exhaustive search. Real evolution is not at all ergonomic, since the space of all possible designs is much too immense for exhaustive search.

It turns out that with this same model there is actually a much quicker ideal evolutionary pathway that achieves fitness $BB(N)$ in time of the order of $N$. This path is however unstable under random mutations, plus it is much too good: Each organism adds only a single bit to the preceding organism, and immediately achieves near optimal fitness for an organism of its size, which doesn’t seem to at all reflect the haphazard, frozen-accident nature of what actually happens in biological evolution.12
So that is the current state of metabiology: a field with some promise, but not much actual content at the present time. The particular details of our current model are not too important. Some kind of mutating software model should work, should exhibit some kind of basic biological features. The challenge is to identify such a model, to characterize its behavior statistically, and to prove that it does what is required.

References


Endnotes
* Chaitin is emeritus researcher at the IBM Watson Research Center, an honorary professor at the University of Buenos Aires, and a member of the Académie Internationale de Philosophie des Sciences. His email address is gjchaitin@gmail.com.

1. **Historical Note:** Algorithmic information theory was first proposed in the 1960s by R. Solomonoff, A. N. Kolmogorov, and G. J. Chaitin. Solomonoff and Chaitin considered this toy model of the scientific method, and Kolmogorov and Chaitin proposed defining randomness as algorithmic incompressibility.

2. **Historical Note:** The idea of running through all possible proofs, of creativity by mechanically trying all possible combinations, can be traced back through Leibniz to Ramon Lull in the 1200s.

3. See the chapter on “The Search for the Perfect Language” in Chaitin, Mathematics, Complexity and Philosophy, in press.

4. **Farewell to Reason** is the title of a book by Paul Feyerabend, a wonderfully provocative philosopher. We borrow his title here for dramatic effect, but he does not discuss Ω in this book or any of his other works.

5. The term *quasi-empirical* is due to the philosopher Imre Lakatos, a friend of Feyerabend. For more on this school, including the original article by Lakatos, see the collection of quasi-empirical philosophy of math papers edited by Thomas Tymoczko, New Directions in the Philosophy of Mathematics.


7. See Neil Shubin, Your Inner Fish: A Journey into the 3.5-Billion-Year History of the Human Body.

8. In order to avoid getting stuck on a local maximum, in order to keep evolution from stopping, we stipulate that there is a non-zero probability to go from any organism to any other organism, and −log, of the probability of mutating from A to B defines an important concept, the mutation distance, which is measured in bits.

9. **Alternative formulations:** The organism calculates a total function \( f(n) \) of a single non-negative integer \( n \) and \( f(n) \) is fitted than \( g(n) \) if \( f(n)/g(n) \to \infty \) as \( n \to \infty \). Or the organism calculates a (constructive) Cantor ordinal number and the larger the ordinal, the fitter the organism.

10. Consider \( BB(N) \) defined to be the maximum run-time of any program that halts that is less than or equal to \( N \) bits in size.

11. Note that to actually simulate our model an oracle for the halting problem would have to be employed to avoid organisms that have no fitness because they never calculate a positive integer. This also explains how the fitness can grow faster than any computable function. In our evolution model, implicit use is being made of an oracle for the halting problem, which answers questions whose answers cannot be computed by any algorithmic process.

12. The \( N \)th organism in this ideal evolutionary pathway is essentially just the first \( N \) bits of the numerical value of the halting probability \( Ω \). Can you figure out how to compute \( BB(N) \) from this?

13. For instance, will some kind of hierarchical structure emerge? Large human software projects are always written that way.

Machines among Us: Minds and the Engineering of Control Systems

Ricardo Sanz
UPM Autonomous Systems Laboratory

I Machines and their Minds
Some years ago my then eight-year-old son asked me: Dad, what is your work about? For a control engineer like myself this is not an easily answerable question (at least for untrained audiences). While controllers are pervasive technologies in the world of today—e.g., they are critical components of our cars, microwave ovens, heart pacemakers, or electrical grids, etc.—the perception of their reality, constitution, and role by the common mortal is not an easy task.

People can see the electronic parts of our surrounding devices or the full- anded computers that sustain complex systems operations; but controllers themselves, being reified abstract signal structures, are not easily perceivable. For the untrained eye, seeing a controller operation inside a machine is as difficult as seeing the Spanish lessons operating when someone asks for “paella” in a restaurant. For sure there is something “Spanish” inside that generates this asking but... what is it? What is its nature, its parts, and its relations with the rest of the body? Is it just something that the brain does? What is the difference between learning to speak Spanish and learning to dance mambo? These are complex issues.

The answer I gave my son—that later has served me as a guiding light for my research—was quite simple, however: control engineers make minds for machines. This is a very simple answer targeting at the not so simple problem of mind construction. The controllers inside the machines are sometimes as elusive as their natural counterparts. However, in the same sense as we can see the brain as the platform for the mind, we can see the machine embedded computers as the platforms for the artificial minds. The computer-machine relation is, in a very precise sense, a mind-body relation.

The purpose of this article is not to raise the flag of computationalism in cognitive science—the old “mind/brain is computational” argument—but to look at the very same problem from just the other side, from the side of the artificiality. In this article I want to address two main questions: “To what extent do controlled systems need a mind?” and “What kinds of minds do they need?”

Philosophy is perhaps the first and last redoubt of thought. It was Aristotle who said that man is the rational animal. He said so guided by an insight of the time: people can do sums.
The machine surpassed me.

But, as Russell said, considering the coming of the calculating machine, “As arithmetic has grown easier, it has come to be less respected.” Computing square roots are not portentous feats anymore. Almost nobody respects today the intellectual powers of machines.

I’m here to vindicate them. Machines may think. Machines may have desires and plans. Machines may love and hate. Machines may feel. Machines, in a word, will let us reach the old dream of philosophy and science: know what is knowing, thinking, wisdom, and, eventually, mind.

2 About Control Systems and Minds

Karl Åstrom, an automatic control professor in Lund, Sweden, once labeled control technology as “the hidden technology.” This is so due to the abstract nature of its realizations that hamper its visibility. It is hidden from view but so pervasive in our current world that it is difficult to believe the generalized ignorance about it even in the most respected academic world.

Controllers are intrinsic parts of most of the machines around us. These are the parts governing their behavior to make them serve us well by dealing with the disturbances that may affect the services they are providing. From the simple operational principles of room temperature thermostats to the complex strategies for automatic take-off and landing of the automated flight controllers of modern planes, controllers range all the spectrum of persistent and reliable machine behaviors.

Controllers being abstractions notwithstanding, they need a physical realization to have the necessary physical causal powers over the machine components. Current control technology is mainly based in the implementation of computer-based hardware/software systems (see Figure 1). An abstract controller is realized in the form of software that is run over a computer and connected to the controlled system—the plant—by means of adequate sensors and actuators.

Figure 1. The controller and the controlled plant—a computer and a car engine in this figure—are the artificial counterparts of the philosophical mind/body dichotomy.

In a sense, we are equating minds and control systems (as has already been widely accepted both in the natural and the artificial domains). The physical controller and the controlled plant are the artificial counterparts of the philosophical dichotomy concerning mind/body relations. This control stance for philosophy of mind offers a vantage viewpoint for the analysis of mental phenomena (somewhat leveraged by the dynamical systems approaches in cognitive science). But in this article we will not focus on control-theoretical analyses of natural mental phenomena but in the description of an evolutionary genealogy of purely artificial controllers that, perhaps not so surprisingly, can be mapped into the genealogy of mental architecture of animals.

3 Cybernetics and control

Today, the science and technology of computer-based control is facing an enormous challenge when increased levels of autonomy and resilience are required from the machines that are providing core services like energy, water, or communications. We can say, without any doubt, that control systems complexity is boosting to levels that are touching the borders of understandability.

We can describe two paradigmatic examples of the challenges that intelligent controllers are addressing today: electrical networks and cars.

The control of electrical production and distribution networks—the system in charge of continuously providing energy in our wall sockets—is a complex collection of tightly coupled physical systems controlled by a heterogeneous collection of semi-autonomous controllers. The control of power plants, the production-consumption balance, the dependability of the provision of the service, the local and global stability of the network—e.g., to avoid propagation of blackouts—are some examples of the major problems that must be addressed in real-time by collections of networked controllers. To augment the problems, some of the subsystems of the network are under the rule of different owners—people, companies, government—hence adding a normativity problem to the physical or epistemological problem addressed by the controllers.

Modern cars incorporate enormous amounts of computing power in the form of sophisticated electronic control units (ECUs) to provide the required control services needed for engine operation, fuel economy, car body stability, anti-skating, anti-block systems for wheels, driver attention monitoring, road condition monitoring, automatic parking, etc. These computers and their associated control software constitute complex networked structures that are in a continuous process of perception and action.

These two are examples of the class of cybernetic appliance that are in charge of controlling some of the engineering resources that are critical for our way of life. What was simple, homeostatic control of certain magnitudes for the controllers of the past (e.g., controlling rotational speed in electrical generators) has evolved into control problems that are affected by uncertainty, distribution, coupling, etc. This amount to a more open control problem where controllers require increased levels of intelligence and cognitive flexibility to overcome these difficulties.

4 A genealogy of controllers

We can say, in an easy talk, that controllers constitute the minds of the machines. Dennett talked about kinds of minds—
Skinnerian, Popperian, Gregorian—in the natural realm and we may wonder to what extent the controller-as-mind metaphor is more than just a metaphor.

The historical trajectory of control systems has produced a complete genealogy of mechanical minds. This genealogy has evolved in parallel with that of artificial intelligence systems but it is, however, much more close to our very conceptions of mind than the classic logic-based AI paradigm. This evolution has been driven by the need of the real-world artificial systems to overcome perturbation, failure, and change and still be viable.

The most elementary control system is an open loop controller, where a device acts upon the physical thing that we are interested in controlling its behavior. The actions are based on a pre-established plan that is blindly followed no matter what happens. When something goes wrong, the controller can in fact make it worse (e.g., consider the way in which washing machine controllers worked in the ‘80s). For most control engineers, open loop controllers are not very interesting entities, but concerning philosophy of mind they are degenerate cases of what can be considered the essence of mind: a continuous process of embedded perceive-think-act.

In this last sense, the controller depicted in Figure 2 is the entry point in the world of the artificial mind. The L0 controller consists in a pair of coupled subsystems: a sensor and an actuator. This may be also considered a degenerate case of an artificial mind (the “think” aspect seems to be missing) but, in fact, the thought is distributed between the two extant subsystems. This L0 controller is also not very good in handling the unexpected. It is prepared to handle a limited degree of variety concerning a sensed magnitude. The actions are totally driven by the sensed information and when this goes out of the pre-specified tolerance band, the behavior of the controller will be not according to specifications.

The next taxon in this genealogy of controllers appears when the controller is able to decide about the current circumstances and filter the incoming information to avoid using established action patterns in wrong conditions. This class of controller is able to pursue changing goals thanks to its behavioral flexibility (see L1 controller in Figure 3).

The question of the origin of goals—if they are intrinsic or extrinsic in animals or in machines—is not going to be discussed here, but a brief comment about the natural/artificial dichotomy seems necessary: it is the position of this author that there is no ontological difference concerning these issues between natural and artificial systems. It is indeed the purpose of this paper to show the closeness between architectures of minds in the natural and artificial worlds.

The next evolutionary state of control systems evolution is the incorporation of mechanisms of memory. Based on them the controller can behave differently in the same sensed circumstances based on memories of the past. This memory mechanism is realized by means of incorporating infrastructure to keep an internal, continuously changing state of the controller. This is, obviously, a state of mind. The L2 controller shown in Figure 4 uses this state at the upper level of the nested control hierarchy of the L1 controller.

Figure 2. The L0 controller; an entry point into the world of artificial minds.

![Figure 2](image)

Figure 3. The L1 controller has the necessary behavioral flexibility to avoid certain classes of inappropriate behavior or to accept changes of goals.

![Figure 3](image)

Figure 4. The L2 controller incorporates an internal state to be able to remember the past and, based on this, better behave in the future.

![Figure 4](image)

Figure 5. The L3 controller is composed by nesting two L2 stateful controllers at different levels of abstraction.

![Figure 5](image)
The stateful controller can change its I/O behavior based on what it has been experimenting in the past. This offers two very interesting features: i) the capability of adequately responding to dynamical aspects of the world (e.g., the system will be able to see increases in sensed magnitudes) and ii) the capability of learning, i.e., changing its internal state to better behave in the future.

Most control systems used today follow this relatively simple structure, especially those based on software programs running on computers, where the state of the controller is captured in the RAM memory of the computer. This is so important that it shall be considered the cornerstone of the massively deployed control technology of today (from thermostats to cardiac pacemakers and DVD players).

The denominations used for the controllers in this taxonomy (L0, L1, ...) is just a systematic way of naming and does not presuppose an evolutionary or derivative order (but in most cases this is the case).

The L3 controller depicted in Figure 5 is somewhat sideways the mainline. It shows a general strategy for augmenting the competences of a controller: the layering of another one upon it.

L3 uses one L2 over another L2 to offer dynamical control and learning at two different levels of abstraction or focus. This implies sensing different magnitudes (e.g., derived magnitudes at the lower level), performing level-different actions (e.g., setting goals), and managing states of different nature (e.g., physical magnitudes and economic ones).

The L4 controller shown in Figure 6 enlarges and completes the inner state as to be a representation—abstract, partial, approximate—of the world of the control system (the physical plant and its environment). The main difference with the L2 controller is the existence of a mapping—a mathematical morphism—the sensed world and the model used by the agent in its control. While the L2 state let the control agent handle world dynamics in the actions performed by the agent—having a form of anticipation—the existence of the model enables not only this but also the use of the model for other functions (e.g., postdiction or diagnosis).

The existence of the model, as independent of the controller execution and actuator, enables the targeting of certain business-specific magnitudes, opening a new world of optimizing controllers. In the world of large-scale process control systems—oil, gas, chemical, food, etc.—it is said that “controllers make the product; model-based controllers make the money.” From a philosophical standpoint this addresses the core epistemic/ontic issues of the questions of intentionality.

The L5 controller shown in Figure 7 shows another strategy for enlarging the operational competences of controllers: integration. The layering of the L3 controller can be done atop a collection of lower level controllers leading to an integrated controller that addresses a single goal set at a higher level. Lower level controllers are usually controlling specific parts of the system, being distributed across it. Distributed control systems are a core technology today (that underlies both the electrical network and the car examples described earlier). In other cases, the reason for having several low level controllers is not physical distribution but separation of control concerns.
(e.g., when a controller is controlling pressure and another one is controlling temperature in a chemical reactor).

Figure 8 shows an integrated controller where the upper layer used a model-based structure but it could also be any other class of controller.

**Figure 9.** The L7 controller includes metaobservation and metacontrol capabilities, touching the very essence of self and self-awareness.

The next step in this taxonomy appears when the control knowledge used in realizing the control executor is elicited from it and made explicit. The executor is then replaced by a pair engine/knowledge—borrowing the terms used in many knowledge-based systems, e.g., expert systems. The engine captures the general methods for applying control knowledge. The knowledge base captures the specific knowledge needed for this particular controller (i.e., for achieving a concrete set of goals over a concrete physical plant).

The main advantage of knowledge-based controllers is the possibility of incorporating knowledge in very heterogeneous forms—e.g., extracted from human experts’ interviews or based on first principles—and the increased visibility, understandability, and transportability of it.

The next step, shown in Figure 9, is a critical, ongoing step in modern control technology and, at the same time, sits at the forefront of cognitive science, philosophy of mind, robotics, and AI. The L7 controller shown in this figure is enlarged from previous controllers by the incorporation of a nested layer (like the L3/L4 controllers) that has a special property: the domain of its control is not the physical thing underlying the whole controller; its control domain is the lower level controller itself. In a sense, considering the control system as a whole, the controller is observing itself, having a theory of itself and acting upon itself based on this perception and theory.

The reasons for applying this sophisticated control architecture to technical systems is to be able to tolerate uncertainties that go beyond the environment and the physical system and sit at the very controller itself. The need for clarification and precision of the issues involved in this class of architecture—a necessary condition for rigorous engineering—can help clarify some of the core problems in the philosophy of mind and psychology of consciousness.

The final taxon in this taxonomy is shown in Figure 10. The L8 controller is an L7 controller enlarged with a dialogue system including explanatory facilities. Due to the complexity of the L7 controller and the general requirement in complex control systems concerning having always humans on top, able to observe system operation, the system is required to explain itself. This can be done thanks to the availability of a metamodel that contains the meta-information of relevance.

**5 Conclusions**

Computer control systems have been merging with some of the artificial intelligence technologies in a desperate search for task-oriented, robust information processing competencies. But this is touching the barrier of complexity. In a sense, software...
intensive controllers are becoming too complex to be built by traditional software engineering methods, this complexity appearing as an increase in size and a decrease of system dependability; this last being the most frightening one. Apparently there is only one strategy ahead: make the systems take care of themselves. Systems must be more aware and more responsible for themselves and the services that they are providing than they are today. Machine consciousness—consciousness being the central tenet for mind form many of us—is now in the agendas of engineers.

A question emerges: Can artificial consciousness be a solution to the performance/dependability problem in technical systems? Another question also emerges: Can this technological approach give clues to the understanding of natural minds? To our understanding, building this class of controllers and analyzing their behavior can help solve the age-old conundrums about knowledge, mind, self-awareness, or even qualia.

References

Endnotes

—— Philosophy and Computers ——

Towards a Distributed Computation Model of Extended Cognition*

Thomas W. Polger
University of Cincinnati

1. The Pre-History of Extended Cognition
In the early years of the 1990s, a number of philosophers and cognitive scientists became enthused about the idea that mental states are spatially and temporally distributed in the brain, and that this has significant consequences for philosophy of mind.

Daniel Dennett (1991), for example, appealed to the spatial and temporal distribution of cognitive processes in the brain in order to argue that there is no unified place where or time when consciousness occurs in the brain. Dennett used this interim conclusion as a premise in a series of arguments designed to undermine what he calls the “Cartesian Theater” view of consciousness, according to which there is an independent and fact of the matter about when and where conscious mental states occur. Most famously, Dennett argues that anyone who believes there are such facts of the matter must choose between “Orwellian” and “Stalinesque” theories of the timing of consciousness, in order to accommodate certain data about the spatial distribution of the timing of consciousness-related brain events. But, he claims, that distinction is a distinction without a difference: The choice between “Orwellian” and “Stalinesque” views is either nonsense, or at the very least one that we could never be justified in making. And this, Dennett takes it, is a reductio ad absurdum of any theory of consciousness that says there is a fact of the matter about when and where it occurs.

A second early advocate was Andy Clark (1996) who, inspired in part by roboticist Rodney Brooks (e.g., 1999), argued that much of our cognitive economy runs on piecemeal sensory and motor processes that are spatially distributed throughout the body, and is “scaffolded” by spatial and temporal relations that hold between the body and the world. Clark and Brooks took this psychological and engineering result to show that cognition does not require the kinds of unified and localized representational structures and cognitive modules that are the stock and trade of cognitive science, even in its connectionist forms. Like Dennett, Clark and Brooks hold that there is no single and total fact about the representational state of a cognitive system at any particular moment, or any single place where or time when the cognitive processing occurs.

In subsequent years, enthusiasm for these particular appeals to spatially and temporally distributed phenomena has faded. For one thing, despite a recent flurry of interest in the timing of the “feeling” of free will (Wegner 2002), it was widely appreciated that the timing phenomena cited by Dennett could be given alternative explanations. And, with respect to the “non-representational” cognition suggested by Brooks, both he and Clark acknowledge that it would be more accurate to say that the systems lack central representations, rather than that they lack representations at all (Brooks 2002). So by the late 1990s there was sufficient reason to doubt that the spatial and temporal distribution of cognitive systems was as radical, or had as radical consequences, as its early proponents had claimed.

Furthermore, we might add, the very claim that cognitive systems, processes, or representations are spatially and temporally distributed was never subjected to careful analysis. It is not at all clear, for example, that paradigm cases of distributed systems, such as distributed computing systems, have the radical features that were supposed to go along with such distribution.

2. Embodied and Extended Cognition
Lack of enthusiasm for a few of the more radical arguments advanced by Dennett, Clark, or Brooks should not be understood as indicating that there is no interest in the spatially and temporally distributed processes that attracted their attention. Recently the focus has turned away from the ways that cognitive processes or states are spatially and temporally distributed within systems to the ways that cognitive states and processes may be spatially and temporally extended between systems, or between systems and their environments.

According to advocates of “embodied” cognition, including Clark and Brooks, cognitive processes constitutively involve “non-cognitive” aspects of the system in which (or of which) cognition occurs. These non-cognitive aspects might include the morphology of the system, as well as facts about the system’s ability to interact with its environment that constrain cognition and action. Such constraints may modulate, simplify, and sometimes eliminate cognitive burdens that would have to be borne by the cognizing system according to traditional cognitive science. For example, a system might not have to calculate and initiate compensatory actions to maintain its balance, if instead the musculature of the system automatically makes
such compensations.

And according to advocates of “extended” cognition, cognitive processes can constitutively involve distal objects and processes, or relations to them. That is, cognitive processes can constitutively involve processes that are external to the system’s container, the “skin-bag” as Clark calls it in the case of human beings (2004). It is extended cognition that will be our topic henceforth, so it will be worthwhile to get clear about its core idea.

The canonical formulation comes from Clark’s paper with David Chalmers, “The Extended Mind” (1998). There they argue that cognitive systems can become coupled with elements of their environments in ways that allow those external bits of the world to be actively involved in the cognitive processing itself—not just as objects or stimuli, but as genuine gears of the cognitive engine. Clark and Chalmers tell the story of two characters, Inga and Otto, who wish to find the Museum of Modern Art (MoMA) in New York City. Inga remembers that the MoMA is on 53rd Street, and she is able to use that memory to guide her action. Otto, however, suffers from a memory disability that does not allow him to recall the location of the MoMA in the same manner that Inga does. Fortunately, Otto is aware of his condition and plans accordingly: he always carries a notebook with him in which he writes down the information that he will need later, and he is able to quickly, reliably, and accurately make use of that information. So while Inga consults her memory, Otto consults his notebook.

But what is the difference between Inga’s neurons and Otto’s marks on paper? Both can be quickly, automatically, constantly, and reliably drawn upon to guide action. Indeed, Otto’s notebook could be even more constant and reliable than Inga’s memory:

In these cases, the human organism is linked with an external entity in a two-way interaction, creating a coupled system that can be seen as a cognitive system in its own right. All the components in the system play an active causal role, and they jointly govern behavior in the same sort of way that cognition usually does. If we remove the external components the system’s behavioral competence will drop, just as it would if we removed part of its brain. Our thesis is that this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head. (Clark and Chalmers 1998: 8–9, emphasis added).

According to Clark and Chalmers, Otto’s notebook can be coupled to him so that it is every bit as much a part of his cognitive system as Inga’s memory-laden neurons are part of her. The only difference is that Inga’s cognitive processes are contained within her body, whereas Otto’s cognitive processes “loop” into and are distributed between him and the world—they are extended cognitive processes (Figure 1).

The extended cognition thesis has been widely elaborated since Clark and Chalmers introduced us to Inga and Otto. (And it has been widely criticized.) It has been frequently noted that the thesis depends on a broadly functionalist or computational theory of mind, and a rather abstract version no less. And Fred Adams and Ken Aizawa argue that the extended cognition thesis depends on a fallacy, confusing causal coupling with constitution (e.g., 2001, 2008, 2009a, 2009b). Coupling, they argue, is perhaps necessary but not sufficient for constitution. For one thing, causal-functional coupling between cognitive/computational systems and their environments is abundant, indeed ubiquitous. Every creature is coupled with its environment in indefinitely many ways. But the thesis that cognitive processes extend outside the body and into the environment would be trivialized if every reliable causal interaction was a potential constitutive coupling. The interest of the thesis lies in the idea that there are actual but nevertheless relatively scarce kinds of cognitive coupling that allow extra-bodily processes to become constituents of cognitive processes. The question, then, is what scarce sorts of coupling are sufficient for cognitive constitution?

3. Back to Distribution

To answer the question of which sorts of coupling are sufficient for cognitive constitution, it will be useful to take a lesson from the early advocates of “distributed” cognitive systems. One trouble, recall, was that the fact that a system is spatially and temporally distributed does not by itself guarantee that there is no fact of the matter about what representational or computational state it is in at any particular moment. To put it in a slogan: Not all causal distribution is computational distribution. If cognition is a kind of computational or representational process, then it seems that spatially extending cognitive systems or processes is not sufficient for cognitively extending them. So if we’re going to try to extend cognition by in some special way coupling spatially distributed components, not just any kind of coupling will do. The kinds of causal coupling that advocates of extended cognition need will be those that are cases of computational coupling. That is, we can offer them the view that physically coupled but spatially or temporally extended components are jointly constitutive of a cognitive system just in case the cognitive process they implement is itself computationally distributed/extended across them. What we need to find is systems that are not merely causally, spatially, and temporally distributed but also ipso facto computationally distributed.

This is a nice proposal, but now we need to know what kinds of causal, spatial, and temporal distribution or extension are ipso facto computational distributions or extensions? Here’s a suggestion: Let us suppose that a computational process will be distributed or extended if it operates on representations that are themselves distributed or extended. Why think this is a good approach? If the representations are not themselves extended, then it may be more plausible to say that the externally located representations are inputs or outputs of the system, rather than saying that they actively participate in the system. This is a natural response to the example of Otto’s notebook: Representationally, Otto’s notebook is fully self-contained. Despite Clark’s and Chalmers’ claims that the example illustrates “active” extended cognition (in contrast to the “passive” extension of Putnam/Burge-style wide content), the case is not so clear. Otto’s notebook, though it is manipulated by Otto, is not manipulated by his cognitive system (a part of Otto) as a constituent of its control of Otto. Rather, relative to Otto’s cognitive system, the notebook is a passive piece of information or environment, a tool that can be taken up or not. Its representational value and content do not constantly depend

Figure 1. The looping model of extended cognition.

---

— 18 —
on the agent that make use of it, or on the fact that it is coupled to him. The content is not itself representationally extended between the subject and the notebook.

A stronger case for extended cognition could be made if the cognitive system makes use of representations that are themselves constitutively extended into the external environment, and not fully contained either within or outside the system. Instead of cognitive processes that “loop” out into the world to pick up information, we would then have cognitive processes that make use of information that is already extended between the thinker and the world, not just available in the world to be accessed by the thinker (Figure 2).

At last we have come to the crux of the matter: What is it for a representation to be representationally distributed? We need a kind of representation that is not merely spatially or temporally distributed, but that is also and ipso facto representationally distributed.

**Figure 2.** The distributed representation model of extended cognition.

### 4. Distribution and ‘Real’

What is it for a representation to be representationally distributed? There are at least two good candidate answers to that question:

(i) Representations are distributed when their vehicle pieces are not individually semantically evaluable and syntactically structured.

(ii) Representations are distributed when their vehicle pieces are individually semantically evaluable and syntactically structured.

How could these contradictory theses be options for understanding distributed representation?

Here some examples will be useful, so consider the representations of the number 173 that are illustrated in Figure 3. Most of these are standard representations in various numeral systems; let (h) be stipulated to be a representation of 173 for our purposes. Each of the examples (a)-(h) differ with respect to their spatial extension, and perhaps temporal extension as well. For example, (c) is plainly more spatially distributed than (b) or (a). And (a) and (h) appear to be most spatially compact, least distributed. (But if only the inked area is part of the representation proper, then (a) and several others may be smaller and less distributed representations than (h).) So Figure 3 illustrates a variety of representations of the same content that vary in their spatial and perhaps temporal distribution.

Which of these representations is also (most) representationally distributed?

Right away we can notice that mere spatial and temporal distribution does not make for interesting representational distribution. Example (c) is more spatially extended than (b) and (a), and there is a plain sense in which its representational capacities are more distributed than theirs. Moreover, that distribution has practical consequences: some systems will take longer to process the information encoded in (c) than that in (a); using them may require different perceptual or storage systems, or at least the efficiency may be modulated by them. Yet it’s fairly clear that the difference between (a), (b), and (c) is not what has usually interested philosophers and cognitive scientists. That difference, for example, in no way suggests that there is no determinate fact of the matter about what value is represented by (c), or that it is any less determinate than (a). We will return to this line of thought in a moment. But first let us turn our attention back to options (i) and (ii).

On the one hand, we might think that a representation is representationally distributed in case its representational significance—its ability to function as a representation, and as a particular representation—is relatively immune to changes in its representational vehicle. In this case, example (g) seems to fit the bill. The prisoner’s tally (unary) representation is robust in that it continues to be able to represent a quantity regardless of modifications to the vehicle, and the quantity or number represented approximates the original value in proportion to the modification to the vehicle. Not so for (a)-(c) where small changes to the vehicle can create large changes to the content. Or for the unstructured (h), whose content might be changed arbitrarily or eliminated by any changes to the vehicle. (Perhaps it is the case that a similar symbol that lacks the “leaf” element represents aardvarks.) So there is a good sense in which (g) is representationally distributed, and indeed the most so of the examples in Figure 3.

| (A) 173 | (E) CLXXIII |
| (B) 173 | (F) 10101101 |
| (C) 173 | (G) |
| (D) one hundred seventy-three | |

On the other hand, every element within (g) has representational content independently of its participation in the whole representation. So the representational content of the spatially distributed (g) is merely a function of its composition out of elements whose representational capacity is (relatively) local and undistributed. In contrast, the representational capacity of (h) requires the presence and arrangement of all of its vehicle elements. And the representational capacity of (a)-(c) also requires the presence and organization of all the elements, or at least it is a non-summing function of the representational content of each element. (Otherwise (a)-(c) might all represent 11, rather than 173.) Insofar as the ability of (a)-(c) or (h) to represent 173 depends on the presence of all...
of their vehicle elements and is not a summing function of the representational contents of those elements taken individually, there is a case to be made that the total representational capacity and content is distributed across all of the elements of the vehicle taken together.7

So there is a reasonable case to be made that either of the contradictory options, (i) and (ii), could be the right way to think about distributed representations. What are we to do? Here we can take a page from J. L. Austin. Though Austin was concerned with the ordinary meanings of words, his analytic tools can be adapted to say something useful about the technical concepts with which we are wrestling. In Sense and Sensibilia (1962) Austin argues that understanding the meaning of the word ‘real’ depends on a number of other factors, on what the subject is and what the speaker is trying to convey. So, for example, ‘real’ in “real duck” (real versus decoy) has a different significance than it does in “real oasis” (real versus mirage or hallucination). Austin says that real is “substantive hungry” in that its meaning depends on the term (object) to which it is applied. Likewise, he thinks that ‘real’ depends on what it is contrasted with, so it is a contrast notion.8 So the meaning of ‘real’ in “real duck” depends on the contrast being decoy ducks say, rather than robotic ducks. Sadly, we can readily imagine a future world in which someone says, “Those aren’t real ducks, those are engineered. All the real ducks are in zoos.” Austin also says of real that it is a dimension term, that it and its negation pick out the most general of a class. From more general to more specific: unreal, artificial, decoy, wooden. Finally, he says that it is an adjuster term, that it allows us to say new things without introducing wholly new vocabulary, by adjusting the meaning of existing terms. We didn’t have to introduce the term ‘decoy’ because we could readily say that some ducks are not real.

The suggestion, then, is that to answer the question of what kinds of representation are representationally distributed in space and time, we shall have to consider the ways in which distributed representations are supposed to modify and contrast with other kinds of representations. Plainly the notion of distribution here is used to adjust our standard notions of representation, the ones to which distributed representation— and, if I am right, distributed or “extended” cognition—are to contrast.

What about the substantive-hungry, contrast, and dimension aspects of distribution? Distributed representations are not just distributed ink splotches—this is why it seems right to say that (a)-(c) are not increasing representationally distributed even if they are increasingly spatially distributed. This is why when the substantive whose distribution we are considering is representation—mere spatial or temporal extension is not sufficient. (Whether it is even necessary is unclear.) Likewise, we shall have to carefully attend to the kinds of representations to which distributed representations are supposed to contrast, and that lie on the axis of which distribution is one terminus.

At this point we know more about what kinds of representation are not distributed. Notice, in particular, that there is nothing special about the representations in Otto’s notebook. They are not distributed: the representations in the notebook are fully encapsulated and local. This, I wager, is why the example of Otto’s notebook is unconvincing to many philosophers and psychologists. It would be more convincing if Otto’s notebook were only part of the representational vehicle—if the representations themselves were extended and non-local. Distributed representations are on one end of an axis of which local representations are the other extreme. This is the contrast to which distributed representations are compared. Of course, the distinction is one of degree, but Otto’s notebook seems to be on the wrong end of the spectrum.

5. Prospects for Extended Cognition

The reader will perhaps be disappointed, though not surprised, that I do not have a full theory of distributed representations to offer. Or even an outline of one. But that is not my goal. Rather, I have been concerned to sketch an alternative approach to extended cognition that does not depend on mere spatial and temporal extension. That, as we have seen, is at best necessary but not sufficient to produce extended computational systems. Actively extended computational systems—and thus actively extended cognitive systems, if cognition is a species of computation—should be systems that manipulate distributed and extended representations (per Figure 2). These will be representations that are already distributed between the core system and the world, not just those that the system can “loop” into the world to access (per Figure 1), however quickly or automatically.

I do not know whether there are any such systems in nature. And I have not argued that Clark, Chalmers, or someone else could not convince us that some “looping” systems are genuine examples of active extended cognition. But I suggest that the distributed representation model is worth exploring.

References


Endnotes

* A version of this paper was presented at the Central Division meeting of the American Philosophical Association in 2010. I am grateful to the audience on that occasion, and to Piotr Boltuc for the invitation to speak.


2. Here I am allowing myself to slide between talk of extended cognitive systems and extended cognitive processes, about which see Adams and Aizawa 2008. The distinction appears to be orthogonal to my present concerns.


4. This is something like the inverse of the “Systems Reply” to John Searle’s “Chinese Room” example (1981). But in this case the complaint is that the notebook is not part of...
the cognitive system because it takes the whole cognitive system to manipulate it; no part of the cognitive system can make use of the notebook without all of Otto making use of the notebook. Not so with memories or beliefs, which can be used by parts of Otto (e.g., by Otto’s associative learning systems) without Otto as a whole taking any action.

5. Perhaps the representations in the notebook depend historically on being produced by someone; but that is not the kind of extended coupling that interests Chalmers and Clark, or other advocates of extended cognition.

6. We can stipulate that it is so. We can make similar stipulations about (g), but only by changing what or how (g) represents 173; that is, by changing a whole representational regime, not just stipulating about a single vehicle.

7. One might think that the elements must be coupled in a particular way. But then we will be running in circles.

8. Austin says it is a “trouser” term.

Figure 1. Part of the related ideas network in the neighborhood of “Turing test.”
face—the InPhO data can be explored via human-friendly HTML, or in a machine-friendly JSON format, and they can be switched between with the simple expedient of adding either .html or .json to the unique URI of each resource. (Actually, the HTML format is the current default, so it can be omitted.) You are invited to explore the API at [http://inpho.cogs.indiana.edu/doc/examples.htm](http://inpho.cogs.indiana.edu/doc/examples.htm).

The power of the approach can be illustrated by example. If interested in philosophical discussion of ideas related to the Turing test a person could look at [http://inpho.cogs.indiana.edu/idea/1039/related](http://inpho.cogs.indiana.edu/idea/1039/related) and follow the links given there. A program (or programmer) can access that same information in a structured way simply by tacking the extension `.json` to the end of the previous URL. Each item in that set of results contains already its list of related ideas, making it relatively easy to build a visual representation of the network of terms involved (e.g., Figure 1).

The original services offered by the InPhO website have now been recoded against our own API which will greatly simplify their maintenance and development in the coming years, as well as enabling us to increase the pace at which new tools can be prototyped and released. These are some tools already in the pipeline:

* An interface that will go beyond the existing “Related Entries” sections of SEP entries, providing an expanded list of suggestions for related topics.

* An interface that will give users access to the bibliographic content of the SEP, to manage collections of bibliographic items, to find these in other philosophical resources online such as PhilPapers, and to export such collections for use in programs such as BibTeX and EndNote, or as preformatted text suitable for import to a paper.

* Access to various alternative ways to visualize the networks of concepts and thinkers represented in the SEP.

* A service that will analyze author-submitted documents and their associated citation lists to suggest other items that the author might want to read.

References
