The Emergence of Self

Terrence W. Deacon, James W. Haag and Jay Ogilvy[1]

Starting small

Rene Descartes' now legendary claim—"I think therefore I am"—sets the challenge for a theory of self. Who or what is this "I" of which Descartes speaks? This is one of those irritating puzzles that perennially re-emerges to challenge philosophers in every age. On the one hand, it is undeniable that the phenomenal experience of being a self is ubiquitous. On the other hand, the nature of this self we experience eludes typical forms of explanation.

In this chapter, we are concerned with the very possibility of explaining the existence of selves. Along this path, our current intellectual ethos typically leads us in one of two directions: We either follow David Hume and disavow any self over and above any set of mental and physical processes (the self is *a useful fiction*) or we emphasize the experience of having a self and assume it to be a brute fact of the world (the self is *a phenomenological experience*). On the first path, which we can identify with eliminativism, the interpretation of our personal experiences as evidence of the existence of a separate subject which has these experiences is called into question. It is an inference from these experiences, not a fact in itself, and so could be mistaken. As Hume reminds us, there is no self to be found separate from these experiences, and so our projection of an entity that contains or possesses these is unjustified. On the second path, which takes its lead from Descartes and we can find articulated by various phenomenological paradigms, comes a focus on the first-person experience of having a self. This view maintains that first-person experiences are both ineffable and undeniably present, and this makes them unquestionably real, and the ground for all other assessments of reality.

We believe that these two options force us into a false choice. This is because both approaches reflect a failure to adequately deal with issues of teleology. Eliminativist approaches deny the reality of teleological relationships, while phenomenological approaches assume it as an unanalyzable primitive. Selves are ultimately defined by their teleological properties. They are

[1] Acknowledgements: The authors were also aided by extensive editorial feedback from Tyrone Cashman, and discussion with Jeremy Sherman, Julie Hui, and Alok Srivastava.

loci of agency directed toward the achievement of ends, they assign value to these consequences, and they must in some sense define an internal / external relationship, implicitly embodying a self-other representation. So the failure to resolve this issue of the origins and efficacy of teleological phenomena guarantees that the concept of self will pose irresolvable dilemmas and consequently remain ambiguous.

So beginning with Descartes' *Cogito* is an ill-advised approach. It assumes what we must ultimately attempt to explain. And yet to deny its reality seems absurd. Subjective experience is both too special and too complex to serve as a starting point. It is too special because the sort of reflective cognition that Descartes accepts as undeniable only became possible after billions of years of evolution. It is not some general ubiquitous quality of things, even if it is our only window into the world. It is too complex because it is the product of an immensely subtle physical process taking place in the astronomically complicated and highly structured chemical-electrical living network that is a human brain. We believe that trying to make sense of something so nearly intractable as a first step is pointless. It is almost certainly one of the main reasons that discussions of self and of subjective experience have produced little progress in understanding. Descartes' question needs to be set aside until we can assess the problems of teleology and self at the simplest level possible, where it may be easier to dissect these issues and potentially build toward the question of subjective self incrementally.

So we will not begin by treating human consciousness as the only relevant exemplar, or as the singular appearance of the property of self in the cosmos. This does not however force us to find traces of self in stones and drops of water. Selves are associated with life. They are not only limited to organisms like humans with complex brains and subjective experiences, and indeed the self experienced by creatures with complex brains is in many ways derivative (or rather emergent from) the self of organism existence. It is not unusual to identify selves throughout the living world, from simple organisms to complicated humans. While these selves certainly have important qualitative and quantitative distinctions, they also share certain core features of what it means to be a self. We believe that much can be gained by exploring self at this more basic level before trying to tackle the problem in its most complex form.

Recognizing that even organisms as simple as bacteria have properties that qualify them as selves, in at least a minimal sense, suggests that self is not just a subjective issue. This allows us to at least temporarily bracket this troublesome attribute from consideration, while exploring

certain more basic attributes. But in setting this issue aside we have not reduced out the most critical issue. Indeed, issues of teleology, agency, and representation, to mention a few, are still in need of explanation, and perhaps unpacking these challenging concepts in simpler contexts can provide clues to the resolution of some of these more complex issues. Nor have we reduced the problem to a merely scientific and physical issue. To the extent that only organisms—and not stones, clouds, streams, or even our most complex computing systems—are selves, it is clear that self is not a simple physical property and not just an issue of complexity alone. It is probably safe to say that 4 billion years ago there was no such thing as self in any form on this planet, and probably not anywhere in our solar system. Physical systems with this property emerged at some point, roughly coincident with the origin of life. The form of self that characterizes human subjectivity is a recent higher order augmentation of this first transition, and so while this complex variant includes such radically different emergent properties as subjectivity and interiority, this phenomenal version of self should nevertheless reflect a common logic that traces to this original transition. We may thus gain a useful perspective on this problem, by stepping back from issues of subjectivity to consider the reasons we describe organisms as maintaining, protecting, and reproducing them*selves*.

The plan of this essay, then, is to first address the philosophical issue of teleology, to offer what we believe is an emergence-based account of the physical basis for true teleological relationships, then to apply this to a basic conception of organism self that addresses many of the component attributes we need to explain, and then finally offer s glimpse of how this way of addressing the issue may help resolve some of the more challenging and personal mysteries of being selves.

The emergence of teleological phenomena

We believe that the primary reason that self poses such a philosophical problem is due to a historical failure to account for the existence of end-directed processes associated with self-behavior. Selves act (or behave) according to a purpose. They have functional components that serve ends and contribute to the integrity of the whole. And they are organized in such a way that achieving or failing to achieve these outcomes has a value. Selves are organized around "final causes" in Aristotle's terminology. Unfortunately, Aristotelian final cause has been treated as an illegitimate explanatory principle in philosophical discourse since the 17th Century

Enlightenment. Philosophers since Spinoza, as we will mention below, have been adamant about the untenable assumptions implicit in teleology. Thus, the useful fiction self and the phenomenological self correspond, respectively, to two dichotomous stances regarding the reality of teleological processes: 1) One can deny teleology in nature and use mechanistic terminology to describe such things as function or design (teleonomic arguments would be an example[2]), or 2) One can assume teleological processes and fail to provide an explanation for their existence and persistence. Are these the only options?

No. We believe that there is a middle ground: a scientific account that can explain how teleological processes in nature emerge from non-teleological antecedents. Although we agree that a direct mapping of phenomenal experience onto physical process is indeed impossible, this is not because of any deep metaphysical incompatibility, but rather because such an account skips over an essential mediating level of complex causal processes. In quasi-Aristotelian terms, we argue that a type of formal causality mediates the emergence of final causality from efficient causality. Instead of trying to reduce final causality to efficient causality (the Aristotelean term for the sort of causality studied in the physical sciences) or showing them to be ultimately incommensurable, we argue that this mediating domain of causal dynamics provides a necessary bridging domain between them. We argue that this intermediate domain of causal dynamics is constituted by processes that spontaneously generate and propagate form—often described as "self-organizing processes." These play a critical mediating role between mechanistic and teleological accounts of causality, by virtue of the way they account for the spontaneous origin of dynamical constraints.

The concept of constraint, besides being a critical concept for defining information, also provides a negative way of defining order. Unlike concepts of order defined with respect to a model or an ideal form, describing a given phenomenon in terms of the constraints that it exhibits delineates form in terms of features *not* exhibited. Concepts of regularity and symmetry thus can be reframed in terms of the redundancy that is inevitable when other degrees of freedom or possible configurations are not expressed. The importance of constraint production, and by implication order production, is its contribution to the intrinsic asymmetry implicit in the notion

[2] See Ernst Mayr, "Teleological and Teleonomic, a New Analysis," *Boston Studies in the Philosophy of Science* 14 (1974): 91-117; Ernest Nagel, "Teleology Revisited: Goal-Directed Processes in Biology," *Journal of Philosophy* 74 (1977): 261-301.

of an end or goal, and the distinguishability of self from other, which is not defined by material properties alone.

Not only does the concept of constraint offer a way to define both structural and dynamical "form," it is the critical determinant of the capacity to do physical work, which the complexity scientist Stuart Kauffman usefully describes as the "constrained release of energy." The capacity to do work, in a physical sense, is critical to another intrinsic feature of self: agency.

We will thus identify self with *the intrinsic constraints that organize the physical work (e.g. of the brain or body of an organism) with respect to functional ends and the requirements of a system that confer this capacity*. To summarize the problem of self in Aristotelean terms, then, we will describe the self as a relationship among formal causes constituting the final causal processes that constitute experience. In this respect, self is effectively a system of self-perpetuating formal causes: a dynamical organization that includes the capacity to continuously maintain or reconstitute that form of organization in the face of intrinsic degradation and extrinsic disturbances.

A contemporary version of the Humean self is developed by philosopher Daniel Dennett. In his assessment, Dennett begins at a place not far from our own: "*Now* there are selves. There was a time, thousands (or millions, or billions) of years ago, when there were none—at least none on this planet. So there has to be—as a matter of logic—a true story to be told about *how there came to be* creature with selves."[3] This approach to establish a sort of "proof of principle" echoes our attempt to find a minimal self. However, there are significant differences in Dennett's efforts as evidenced in his claim that basic biological selves are, "just an abstraction, a principle of organization."[4] While more complex, Dennett's commitment at the organism level to the useful fiction self is echoed at the human level as well. While human selves are "nonminimal *selfy* selves," they remain a theorist's fiction: "Like the biological self, this psychological or narrative self is yet another abstraction, not a thing in the brain, but still a remarkably robust and almost tangible attractor of properties."[5] Of course, if a self is an abstraction, then there must be an interpreter capable of interpreting these phenomena in this way and if that interpreter must

[3] Daniel C. Dennett, *Consciousness Explained* (Boston, MA: Back Bay Books, 1991), 413.
[4] Ibid., 414.
[5] Ibid.

also be a self we are left with a vicious regress. There cannot be such a self. And if my self is no more than the collection of all these experiential episodes, whatever they are, then there is nothing more in addition to them to be a source of causal agency. This move to refer to selves as empty abstractions is rooted in two commitments: 1) any central command center in organisms or a "Cartesian Theater" in the human brain is impossible to locate, and 2) there is no way to account for causal changes enacted by an abstraction. We are in full agreement with the first commitment, but not the second. Despite its apparent problems, a variant of the concept of abstraction may, however, provide a clue to this form of causal influence.

This problem of "abstraction" is deeply rooted in some of the most basic assumptions of Enlightenment metaphysics. In their haste to reject Platonic forms, and to embrace a nominalistic materialism, where general principles and formal properties are only causally relevant when materially embodied in some specific substrate, enlightenment thinkers inadvertently eliminated the possibility of conceiving of a bridge across this ontological gulf. This goes to the heart of the problem in a number of respects. Not only is self unable to be identified with any distinct physical material or energy, neither is the content of the thoughts or experiences of that self. How can what is not present influence what is?

In answer to this quite general criticism, we take a page from information theory. Information, as Claude Shannon[6] defined it in a classic 1949 monograph on the topic, is not something present; not a signal or sign or magnetic orientation of an iron fragment on a computer storage medium. Information is something removed: uncertainty. He demonstrated that information is measured in terms of how some medium used to convey it is constrained from exhibiting states that it could have been in. For example, when in 1775 Paul Revere saw two lanterns shining in the old North Church in Boston instead of one, the uncertainty about British troop movements was eliminated. The 50/50 uncertainty of the day before was reduced by this either/or signal. No choice, no information. In this way information is a relationship to what is not exhibited. When a search party fans out into the woods to locate a lost child, the people who find nothing are contributing as much as the one discovering the child. Constraint refers to options, or degrees of freedom not realized— something not immediately present and not physically intrinsic. But even so, a constraint is something quite precise.

[6] Shannon, C., and Weaver, W. (1949).

While treating self as an abstraction in the sense of a description or comparison leads to the conclusion that self cannot be a source of causal power, treating self as the source of constraint on the physical processes generated by an organism has precisely the attributes we require. To perform work and thus alter the physical state of things requires the constrained release of energy. The enclosure of an explosion by the piston and cylinder of an internal combustion engine or the diversion of a stream by a water wheel constrains the release of the energy of these processes so that it can be directed to achieve a desired physical change, like the movement of a vehicle or the grinding of grain. Constraint is, in this respect, exactly the sort of attribute that should be contributed by a self. If self is an abstraction in this sense—a form imposed upon the energetic processes of the world—it can introduce asymmetric causal properties such as are a necessary defining attribute of end-directedness. But we still need to explain the autonomy of this form: how it arises of itself to become a locus of asymmetrical causal influence.

The Persistence of *Telos*

Consider this phrase: "For Nature, like mind, always does whatever it does for the sake of something, which something is its end."[7] For Aristotle, this statement expresses an ostensibly unproblematic view of reality—one shared by many thinkers throughout history. The goal or *telos* of an action "causes" the instantiation of that very goal. When a carpenter builds a table, we recognize that the table was first represented in the carpenter's mind as an end. Tables are not the result of random, unintentional human actions. The seventeenth century brought with it a shift in worldviews, a move from the organic image of nature to a mechanistic alternative. The Aristotelian perspective dominated thought until the Enlightenment. With the rise of modern science, many of the most influential thinkers of the age questioned its legitimacy (even its possibility). The melding of an atomistic metaphysics with Newtonian science—where causation is thought only to occur as collision-like interactions between very basic particles under determinate laws—collapses Aristotle's causal schema into efficient causation alone. Under this new mechanical philosophy of nature all matter is actual (i.e., nothing is potential), with its only attribute being extension in space. Appealing to the final state of an object is impossible (e.g., the

[7] Aristotle, *On the Soul*, in *The Complete Works of Aristotle*, 2 vols., Jonathan Barnes, ed. (Princeton, N.J.: Princeton University Press, 1998), Book II, Part 4.

chair from the carpenter's concept). As epitomized by Baruch Spinoza: "All final causes are nothing but human inventions."[8] This new worldview, as originally expressed by thinkers such as Descartes, Hobbes, and Boyle, which solidified the opposition to teleological explanations, continues to reign in philosophy and science to this day.

The future is literally "no thing"—how could it possibly be a cause? If science has drilled home any concept, it is that change is a function of the material and energetic features of the immediate contiguous past. In the case of the carpenter's intention, we recognize that it is a mental representation, and not some as yet nonexistent future table that is the cause. But this sort of cause is equally troublesome. What sort of thing is this mental representation? It is not the complex neural state that represents it, and yet without this there would neither be a representation nor the organizing process that guides the carpenter's actions. Isn't the content of the carpenter's thought also an abstraction? Indeed, in the same sense as we considered above, we can say that the content is precisely what is not there, that which constrains the neurological activities that are present. And this means that these constraints are what enable this neural activity to do the work necessary to stimulate the controlled release of metabolic energy and coordinate the resulting muscle movements in the pursuit of this imagined end. Of course the self that we have described as a carpenter is not the neural activity and not even this content, but rather what generates this content. What could constitute the autonomy of this process?

In his effort to make philosophy compatible with the science of his day, Immanuel Kant[9] focused considerable attention on the question of how science should regard teleology. Like many others, Kant recognized that the mechanical explanations for nature seemed to leave something out. Specifically, he found machine analogies to be unsatisfying when biological phenomena are considered. Although Kant was a committed follower of the Newtonian worldview, he knew that to make sense of a purpose or end, it would have to be a *naturzweck*, a natural end.

What status must be reached in order for a thing to be a natural end? Provisionally, Kant sets the minimum requirement: "A thing exists as a natural end *if it is both cause and effect of itself*." To meet this requirement, Kant believes there are two principles that will allow us to

---

[8] Spinoza, *Ethics*, 108.

[9] Kant, Immanuel, *Critique of Judgement, Part Two: Critique of Teleological Judgement,* trans. Meredith, Oxford, The Clarendon Press, 1952, p. 18 (marginal pagination, 371).

establish a distinction between "natural ends" (as in e.g. a living organism) and artificial "ends" (as in e.g. a table). In the first principle, a thing is an end if its parts "are only possible by their relation to the whole." So, without the concept of the table in the carpenter's mind (the end), the legs of the table (the parts) are meaningless. Thus: "It is the product…of an intelligent cause, distinct from the matter, or parts, of the thing, and one whose causality…is determined by its idea of a whole made possible through that idea." However, there is a second principle that moves us beyond the realm of artificial ends to natural ends. Kant writes: "[T]he parts of the thing combine of themselves into the unity of a whole by being reciprocally cause and effect of their form." We have now eliminated the carpenter from the picture and stated that in order for this thing to still qualify as an end, its parts must "combine of themselves." For Kant, this combination is a type of "bildende kraft" ("formative power"). While the table meets the first requirement of qualifying as an end, it fails as a natural end because one leg of the table does not produce another, nor does one table produce other tables. We agree with Kant that this is key to establishing the existence of a natural *telos*.

Self-Organization and Reciprocity

Kant notes that only a living organism is a natural end because it is the only phenomenon in the world that can be described as a "*self-formed being*." Kant states: "...an organized being possesses inherent *formative* power…a self-propagating formative power." With Kant's notion of formative power comes the challenge of explaining how an organism is able to form itself. That is, in distinction from a machine in which there is an outside designer setting up the constraints by which the machine's function is determined, we need a way of having a similar process occur intrinsically, without the designer.

Organization is not the norm in the world. As the second law of thermodynamics tells us, left unattended, everything slips into disorder. We intuitively recognize that increasing organization or just preventing spontaneous disorganization from occurring takes outside effort. My desk doesn't organize itself, I must do the work to make the change. However, there are some physical processes that do spontaneously increase in orderliness over time. These are often described as self-organizing processes, though the invocation of this concept of "self" is potentially misleading in this context, since all that is meant is that the increase in order does not trace directly to any extrinsic cause. Examples of self-organizing processes include whirlpools,

frost polygons, and snow crystals. These sorts of spontaneously regularized processes all, interestingly, are generated in systems under constant perturbation, but where these disturbances compound with one another in such a way as to increasingly correlate with one another. The process of becoming increasingly regular is a process of generating and spreading constraints. In the development of a whirlpool for example, what begins as a disrupted flow of water becomes progressively symmetric in organization as different regions of noncircular flow tend to cancel one another's motions and regions of circular flow reinforce one another. What is important about such processes, from our perspective, is that the regularity develops over time as a result of biases of interaction among a vast many components compounding with one another. In this sense, the regularity that emerges is a function of intrinsic factors expressing themselves as a result of constant external disturbance. So long as the disturbing influence continues, the organizing effect is maintained,

It is not, then, a coincidence that the chemical processes that constitute living organisms are for the most part arranged in ways that produce self-organizing effects. The organism is in a constant state of renewal in which new organization is produced (formed and reformed) that allows it to maintain itself. At every moment an organism's material constitution is different, and yet its structural and dynamical organization remains within narrow variational limits, i.e. its organization is highly constrained. So although no new matter or energy is generated, an organism must continually generate and preserve constraints. In this respect it acts on its own behalf. This minimal persistent "self" that is the beneficiary of this formative process, is not identified with the material or the energy of this process, but with the preserved organization and its capacity to organize work that preserves this capacity. What persists into future generations is not its "stuff" or its energy, but the constraints that constitute the organization of this stuff.

Consider a very simple organism like a bacterium. All parts of this organism are in a continuous state of turnover as it both responds to and resists thermodynamic dissolution while also compensating for a changing environment on which it depends for raw materials and energy. The molecular "parts" of this organism do not even enjoy any kind of existence *as* parts independent of this organization, since each is dependent on the interactions among others. So although the parts constitute the whole, the whole also generates each part.

This reciprocity is the essence of the special twist on the process of self-organization that constitutes an organism. It is not merely a self-organizing process, but a reflexively organized

constellation of self-organizing processes, each of which contributes in some way to the conditions that make the others possible. So although each component self-organizing chemical process of an organism requires a constant introduction of molecules and energy to be able to sustain the generation of regularity, they need each other to generate the constraints that each requires in order to persist. These processes are, as Kant surmised, reciprocally both ends and means for one another, each process generating intrinsic constraints that promote the generation of other intrinsic constraints by other processes. In this way the constraint maintaining-propagating logic of the organism is in a sense a higher-order self-organizing dynamic among component self-organizing processes. It is by virtue of this higher-order stabilization of component constraint generation processes that the global constraints constituting the reciprocity of the whole are not only preserved, but able to be reproduced. Reproduction is, in effect, simply an expression of this reflexively closed form-producing process. In this respect an organism is a means to produce itself as an end.

With this basic understanding, we are now in a position to ask: In what sense is the organism a self? If the organism is continually re-produced via synergistically interacting self-organizing processes, then defining self in substantive terms is problematic. Many self-organizing processes in living organisms are multiply realizable, that is, not limited to a single type of molecule or even any specific chemical reaction. So any search for the essential "stuff" of the organism will inevitably fail. Moreover, the organism is not even any single type of organized process, since these too can change over the course of a lifetime. Instead, the unit of continuity that is the self of an organism is the synergistic relationship between numerous self-organizing processes that constitutes this tendency to preserve this synergy. It is then this special reflexive organization of form- (constraint-) generating processes that determines the closure to formal influences that we recognize as a kind of autonomy. Precisely because organized systems spontaneously tend to degrade, a system that actively regenerates and replaces its components and maintains their interrelationships intrinsically has itself as an end. As Kant suggested, when the end is the means and the means is the end a kind of intrinsic teleology comes into being.

Agency

Stuart Kauffman[10] argues that the defining property of an organism is what he calls autonomous agency. With a bit of unpacking, it is possible to see how this characterizes the kind of recursively organized system we have described above. He describes a system with this property as one that is "capable of acting on its own behalf." This phrase of course already presumes something like a self that acts and benefits from this action, but in the context of the description of organism self that we have been developing this can help us to more precisely analyze these critical features of self: autonomy and agency.

By using the term 'act' he does not simply mean to undergo physical change. An act is goal directed, and it must have the capacity to change prevailing conditions. Additionally, it implies the production of work to initiate or counter some change. An action in this sense is therefore what we can describe as teleological work. As we have argued above, the teleological features of an organism emerge from the synergistic reciprocal closure of its component self-organizing processes. This most fundamental reflexive dynamic has an intrinsic directionality and end that both internal thermodynamic tendencies and extrinsic influences run counter to. It is in this respect that the reciprocity of these component self-organizing processes of an organism can be said to be an act with some function or end.

Something that can be a beneficiary of action is implicitly something that is organized to actively avoid being altered or degraded. A relatively inert physical object resists being altered but does not "act" to defend against this perturbation. A dynamical system with a relatively stable organization may also resist being perturbed, as does a whirlpool or a flame, but although it may change in response to disturbance, we would not want to describe this as acting on behalf of itself. A flame, for example, heats up its substrate to the point where it combusts and liberates more heat to raise the temperature of yet more substrate material. In this respect, a flame behaves in a way that maintains its present dynamical form. It has a self-organizing dynamic. But can we say that the flame behaves in such a way that benefits this form? Eventually, of course, it uses up its substrate and thereby undermines the conditions it depends on. We intuitively do not consider it to be acting or benefiting in any sense because its dynamical organization lacks the reflexivity that we have described for an organism. There is a reciprocity between the action of combustion and the requirements for combustion, but this is with respect to an extrinsic substrate. In other words there is no closure; no circularity of constraints; no means-end reciprocity intrinsic to this

[10] Cf. Stuart Kauffman (1995) *At Home in the Universe.* New York: Oxford University Press.

dynamic. Although life, like combustion, requires utilization of raw material and energy liberated from an extrinsic substrate in order to be able to continue to liberate more in the future, it is now in service of an intrinsic self-maintenant self-propagating dynamic. The dynamics of the flame lacks autonomous, internally reinforced determination of its form, and for this reason lacks a self and cannot be said to either act or benefit, even though it has a self-promoting dynamic.

The philosopher-cognitive scientist Mark Bickhard[11] distinguishes these two forms of dynamic by describing a flame as self-maintenant and an organism as recursively self-maintenant, in the sense of maintaining its self-maintaining logic. Again, as in the case of the term 'self-organization,' the use of the reflexive term 'self' in these contexts does not smuggle in the concept of self as we are trying to explain it, but it does indicate the common circularity of effect that characterizes both sorts of phenomena. What we have shown is that a self, as we have applied it to the dynamics of organisms in general, is organized in a doubly reflexive way: in other words, reflexively organized reflexivity, and recursive recursivity of causality.

Such a system exhibits the property of autonomous agency because it does work to counter intrinsic and extrinsic influences that tend to be disruptive of this autonomy. Its capacity to be a locus of work, and therefore agency, derives from two features of this organization: the capacity to assimilate materials and energy from the surroundings and incorporate them into its reciprocal dynamics and the capacity to generate and maintain dynamical constraints. As noted above, we can describe work as the constrained release of energy. This implies that what specifies different forms of work is not the energy but rather the constraints that channel and organize its expenditure. The reciprocity of the constraints generated by the component self-organizing processes of an organism is in this respect the basis of its autonomy and its agency. By intrinsically generating, maintaining, and reproducing constraints on the flow of material and energy through it, an organism creates the capacity to originate specific forms of work that reflexively reinforce this capacity.

Unpacking the assumption of autonomous agency in this way can help to cast new light on one of the more troubling conundrums of metaphysical philosophy: the problem of free will. Historically this riddle has been posed in terms of a necessary contradiction between the notions of physical determinism and human agency. But as we have defined agency here it is not merely

---

11

causal determination, but rather a specific end-directed form of work. An autonomous self, whether in the form of a bacterium or a reasoning human, is the locus of highly convoluted recursive processes that generate specific forms of work that are organized with respect to some aspect of this autonomous circular dynamics and contrary to some pervasive condition or tendency extrinsic to this autonomous dynamics. This is in no sense contrary to the deterministic cause and effect logic of the physical sciences, but is instead only contrary to some specific local tendency, such as thermodynamic decay. Such tendencies are not deterministic in any strong sense. The Second Law of Thermodynamics, for example, is a tendency—even if it is an astronomically probable tendency—and at least locally it can be countered. Free will can in this respect be recast in terms of a minimally constrained capacity to initiate work aimed at modifying some otherwise prevailing tendency. Thus reduced to its essential features, it can be seen to be entirely homologous with the concept of autonomous agency. As both autonomy and the flexibility to produce more diverse forms of work has increased over the course of biological evolution, so has the relative freedom to interfere with the prevailing conditions of the world in ways that trace their origin to intrinsically generated ends.

From autonomous agency to subjectivity

We have argued above that the core property that links the selves of even the simplest life forms with that seemingly ineffable property that characterizes the human experience of self is a special form of dynamical organization: a doubly reflexive form-generating dynamics. Literally, this is the analogue of self-reference, a logical type violation, and it is not surprising that this feature is even the defining characteristic of reflexive reference in language. Articulating exactly how and why this feature is important for the constitution of a minimal physical self, such as an organism, has helped to unpack many of the assumptions that are implicit in all forms of self: like teleology, autonomy, and agency. It has not, however, provided an account of that most distinctive human attribute of self: its subjective experiential component. Although some might be tempted to ascribe a form of subjectivity to even simple organisms lacking nervous systems, even if this were so (which we doubt), it would only posit the existence of this property by fiat, and would do nothing to explain what difference having a brain contributes. While we argue that there is a common dynamical logic that is fundamental to all phenomena that we consider as having selves, this does not take into account the nested nature of neural dynamics within

organism dynamics, and the additional complication that this complex multi-level multiply reflexive dynamic contributes. Thus, while we have argued that even the simplest bacterium can be said to be organized as a self, and exemplifies the emergence of teleological properties and autonomous agency, it is likely that the subjectivity that we likely share with other species with complex brains involves higher order properties emergent from these higher order reflexive dynamics.

The value of starting small and simple in this analysis is that it has identified what appear to be very general organizational principles that should be relevant to self at whatever level and in whatever form it appears. Before we can apply these principles analogously to the case of human subjective experience and agency, it is necessary to consider what this nested logic of brains within bodies adds to the problem. The relevance of the sort of selfhood that characterizes living organisms, in general, to that more complex form of self that constitutes human subjective experience is made clear by the fact that although the unconsciousness of anesthesia can temporarily interrupt this experience, it can persist across such gaps, so long as the body remains alive and the brain is largely undamaged. Our worries about death, and our comparative unconcern with the state of unconsciousness, is clear evidence that we intuitively judge the self of Descartes' cogito to be subordinate to the self of life in general.

Rather than relying on introspection to provide us with a window on selfhood, agency, or subjectivity, we've chosen to construct an account of self that is based on simpler selves than those of humans. Now that we have that account in hand, we need to consider how the logic of this lower order form of self might point the way toward features of subjectivity and the sense of interiority that is so distinctive of human consciousness.

Given the importance of the doubly reflexive form of dynamics that constitutes organism self, it seems reasonable to expect that something of this logic—instantiated at the higher level that brains provide—is relevant to the account of subjective self. With the evolution of ever more complex forms of organisms the recursive complexity of self has no doubt also grown, but the evolution of brains contributes more than merely a complexification of internal dynamics and of the work that an organism can initiate. It also provides a means to simulate these processes, in service of increasing their effectiveness and flexibility. Since the organism itself, its internal dynamics and its external relations, is also simulated by brains, additional logical type violating loops of dynamics can come into play. On top of this, the capacity for recursive self-reflection

aided by symbolic referential processes that have become uniquely available to humans introduces an even yet more convoluted possibility for reflexive causal relationships. These evolutionary innovations are distinctive rungs on the evolutionary ladder, where the discontinuous emergence of new levels of self, punctuates the spectrum connecting humans to the simplest organisms. Rather than a continuous gray-scale of degrees of self, the evolution of brains and of symbolic communication clearly mark transitions to higher order forms of self dynamics whose constituents are the self-properties of lower levels. So before we turn to the need for any metaphysical magic, it is worth attempting to understand what these further levels of reflexive dynamics might contribute.

The evolutionary framework suggests one further complication. The form of organization we have described as organism self has complexified and differentiated over evolutionary time. The evolutionary appearance of organisms with brains was, however, a special jump in levels, and ushered in an entirely novel emergent realm of self dynamics. This is an important model to also keep in mind in our effort to explain subjective self. Since the function of a brain is in one sense to generate complex neural activity requisite to the complexity of a changing and unpredictable environment and the challenges it poses to the organism, the self dynamics it produces is likely to be as undifferentiated as the dynamics of metabolic maintenance at some times and as differentiated as the complex stimuli and possible interactions required to engage in a complex interaction with other individuals with minds of their own in contexts that are unfamiliar. Indeed, moment-to-moment the level of differentiation of this dynamical synergy must rapidly change, developing from undifferentiated to highly differentiated forms of self in response to changing needs and extrinsic conditions.

The development of one's personal experience of self also has emerged in a process of differentiation. The self that is my entire organism did not just pop into the word fully formed. It began as a minimal undifferentiated zygote; a single cell that multiplied and gave rise to a collection of cells/selves that by interacting progressively differentiated into an embryo a fetus an infant a child and eventually an adult organism. Indeed, it is difficult to imagine subjective self just popping into existence fully differentiated. By the very nature of its thoroughly integrated and hierarchically organized form it would seem to demand a bottom-up differentiation in order to produce it. But if so, then it also suggests that the human subjective sense of self as well is only the final phase in the moment-to-moment differentiation from lower

level less differentiated forms of self-dynamics.

Brains are, after all, organs that evolved to support whole organism functions critical to persistence and reproduction. They are not arbitrary general-purpose information processing devices. Everything about them grows out of and is organized to work in service of the organism. Animal physiology is organized around the maintenance of certain core organism self-functions on which all else depends. Critical variables, such as constant oxygenation, elimination of waste products, availability of nutrients, maintenance of body temperature within a certain range, and so forth, all must be maintained or no other processes are possible. Sensory specializations, motor capabilities, basic drives, learning biases, emotional response patterns, and even rational reflection are ultimately organized with respect to these critical core variables. This suggests that the core undifferentiated form of subjective experience, from which all the more differentiated forms of experience emerge, is organized as are these core organism functions, and serves as a kind of seed from which complex forms of subjectivity differentiate.

So how might these special properties help explain why being an organism with a complex brain includes a form of reflexivity with a mode of reflexive organization that is also reflexively organized with respect to itself? Or to ask this in other terms, what is this locus that "feels" and from which agency not only emerges but to which it is also represented? Again we take our hint from the reflexive dynamical organization that constitutes even the simplest form of self. Since the teleology that distinguishes the agency of organisms from mere work is a product of the closed reciprocity of spontaneous form-generating processes, it is this higher order dynamic that constitutes the self of the organism. Approaching the self-dynamics of brains from the same framework, we would have to say that there must be an analogous closure of dynamical activity with respect to which subjective agency emerges. Without such an origin, the agency of neural processes would inherit its teleological character only from organism self, but if in addition there exists an analogous reciprocal reflexive dynamics generated within the circuitry of the brain itself, there will also be a corresponding neurological source of this teleological orientation, only minimally subordinate to the teleology of the organism. The suggestion is that the subjective self is to be identified with this locus of neurological *telos*; a self-reinforcing reflexive process that serves as a reference dynamic against which all other dynamical tendencies and influences are contrasted as non-self. Though the minimal form of this dynamic may be as undifferentiated as the reciprocally organized metabolic processes that it depends on, its

dynamically facile substrate also predisposes it to differentiate with respect to a complex environment of sensory "perturbations," present and remembered, as well as changing metabolic states.

The supra-individual symbolic tools made available to human brains adds yet a further reflexive loop with respect to which teleological relationships can emerge, and higher forms of agency can be generated. Because we humans can represent our worlds using symbols, which depend on a more abstract logical reciprocity and codetermination, we are capable of forms of work—e.g., the construction of narratives, the creation of obligations, obedience to principles, and so forth—not available to simpler forms of life. Thus it is not unusual for someone to identify with a self-narrative, or "higher purpose" and allow this to become a source of agency. Indeed, we might be tempted to ascribe the agency of supra-human selves like Jehovah or Allah to such a locus of teleology.

In conclusion, we have only briefly gestured toward a new way of understanding subjective self, but the picture of human "selfiness" that emerges from this account is neither Humean nor phenomenological. There is no ghost in the organic machine of the body, because the body is organized as a self of a lower order. There is no inner intender as witness to a Cartesian theater because the locus of perspective is a circular dynamic where ends and means, observing and observed, are incessantly transformed from one to the other. Instead the logic of the mutual reciprocity of constraints creates a relational ontology with respect to which autonomy and agency, and their implicit teleology, can be given a concrete account.

Human subjectivity, when viewed through the perspective of this circular logic of form-generation, is not so much a "hard problem" in the sense of demanding highly sophisticated analytic and scientific tools to solve. It is rather a highly counterintuitive problem, because it requires that we abandon our search for a substantial self in favor of a self that is constituted by constraints, and constraints are not something present, but the boundary conditions determining what is likely. The complex an convoluted dynamical processes we believe to be the defining features of self and any given level are reciprocal limitations on dynamics, not the processes themselves nor the materials and energy that are their instantiation. So ultimately, this view of self shows it to be as nonmaterial as Descartes might have imagined and yet as physical and extended as the hole in the hub of a wheel, without which it would just be a useless disk.

References

Aristotle, *On the Soul*, in *The Complete Works of Aristotle*, 2 vols., Jonathan Barnes, ed. (Princeton, N.J.: Princeton University Press, 1998), Book II, Part 4.

Bickhard, Mark H. (2003). Process and Emergence: Normative Function and Representation. In: J. Seibt (Ed.) *Process Theories: Crossdisciplinary Studies in Dynamic Categories*. Dordrecht: Kluwer Academic, pp. 121-155.

Dennett, Daniel C. *Consciousness Explained* (Boston, MA: Back Bay Books, 1991).

Descartes, René. *The Philosophical Writings of Descartes*, 3 vols. John Cottingham, Robert Stoothoff, Duglad Murdoch, and Anthony Kenny, trans. Cambridge: Cambridge University Press, 1984-1991.

Hume, David. *An Enquiry Concerning Human Understanding.* Amherst, N.Y.: Prometheus Books, 1988.

Kant, Immanuel (1790, 1952) *Critique of Judgement, Part Two: Critique of Teleological Judgement,* trans. James Creed Meredith, Oxford: The Clarendon Press.

Kauffman, Stuart (1995) *At Home in the Universe*. New York: Oxford University Press.

Mayr, Ernst "Teleological and Teleonomic, a New Analysis," *Boston Studies in the Philosophy of Science* 14 (1974): 91-117

Nagel, Ernest "Teleology Revisited: Goal-Directed Processes in Biology," *Journal of Philosophy* 74 (1977): 261-301.

Shannon, C., and Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Spinoza, Baruch. *Ethics*. G.H.R. Parkinson, Trans. and Ed. New York: Oxford University Press, 2000.