Journal of Cognition and Neuroethics

Agency through Autonomy: Self-Producing Systems and the Prospect of Bio-Compatibilism

Derek Jones
University of Evansville

Biography

Derek Jones is Assistant Professor of Philosophy and Director of Cognitive Science at the University of Evansville. He works on topics at the intersection of the philosophy of mind, action and biology and is currently writing a book on the biological foundations of agency.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Jones, Derek. 2015. "Agency through Autonomy: Self-Producing Systems and the Prospect of Bio-Compatibilism." *Journal of Cognition and Neuroethics* 3 (1): 217–228.

Agency through Autonomy: Self-Producing Systems and the Prospect of Bio-Compatibilism

Derek Jones

Abstract
In this paper I motivate a biologically-oriented compatibilism that is consistent with Daniel Dennett's compatibilist account but which avoids some of the recent criticism directed towards it, specifically challenges to his "mild realism" and his reformulation of the principle of alternate possibilities. I argue that a theory of free will that grounds agency in the dynamics of autonomous self-producing systems can show the ways in which agents may engage with and contribute to a given past in uniquely agential ways.

Keywords
Adaptivity, agency, autonomy, autopoiesis, compatibilism, complex systems, Dennett, Jonas, Kant, Lorenz, process biology

Introduction

Determinism is the thesis that past events together with the laws of nature fully determine future events; in a deterministic universe there is exactly one possible future at any given time. One formulation of the problem of free will is that if determinism is true, then our actions are not truly "up to us"—any causally-efficacious state within us would itself have been fully determined by some prior cause, which itself would have been determined, and so on. Given the transitivity of the determination relation, it follows that the initial configuration of the universe, together with its laws, fully determines the final configuration of the physical universe. Anything obtaining between those events is simply along for the ride.

However, there is reason to think that varieties of freedom may nonetheless arise between the birth and heat death of a deterministic universe. Hans Jonas (1966) suggests that such varieties are distinctly biological:

> ... it is in the dark stirrings of primeval organic substance that a principle of freedom shines forth for the first time within the vast necessity of the physical universe—a principle foreign to suns, planets and atoms.

In what follows I sketch how this biological "principle of freedom" may emerge in a deterministic universe. My focus is not on capacities required for morally significant free will, but on the necessary conditions for being a free agent of any kind. It will help to begin with a famous case of naturalistic compatibilism.

Dennett's Compatibilism

In *Freedom Evolves* Daniel Dennett suggests a way in which free will worth wanting might arise in a deterministic universe. He describes creatures in a deterministic "Game of Life"-style toy universe that, upon achieving an appropriate degree of complexity, are best described in behavioral language—they "seek," "avoid," "eat" and so on. These terms describe the systems' *capacities*. To say that a creature in this world *can* avoid harm is to say that it is organized in such a way that it *will* avoid harm in a certain range of conditions. It exercises its capacities for avoidance when it *does* avoid harm. If we were to restart its universe a million times it might avoid harm in precisely the same way each time, but this fact in no way robs the organism of its abilities to avail itself of the opportunities presented by its world (Dennett 81).

Some might not wish to describe Dennett's creature as *freely* avoiding. When we claim that an agent can act freely, we mean that it was possible that they could have done something other than what they did. Genuine freedom involves the agent's ability to collapse a range of possible futures into a single actual event. This notion, known as the principle of alternate possibilities (PAP), has been challenged by Frankfurt (1969) and others, but Dennett accepts it, choosing instead to blunt the challenge by distinguishing between wide and narrow conceptions of possibility.

The narrow reading of "Pat could have Φed" suggests that, given the fixed past up to the time T of the action, Pat could have either Φed or not Φed at T. The wide reading of "Pat could have Φed" suggests that, had the universe been different in some way prior to T, Pat would have Φed at T. Dennett argues that the wide reading underlies most of our empirical tests of causal power. For example, we confirm that one *could* have sunk a missed putt in golf by repeating the putt in circumstances *similar* to those obtaining during the original putt. Performance in *identical* circumstances is irrelevant to the investigation.

This wide reading is entirely consistent with determinism. If what we are saying when we claim that Pat could have Φed is simply that, had previous conditions been different, Pat would have Φed, we are identifying a range of possible deterministic unfoldings of the universe (individuated either by starting conditions or laws) that

happen to include Pat. The agent's causal powers cash out in terms of how competently it copes with whatever unfolding it happens to face. If Pat could only successfully Φ in one or two of the various relevant possible timelines leading up to T, we might judge them as being less competent than an agent that Φs across a broader range. In some cases we might be warranted in chalking the performance up to luck. Competence—what the agent *could* do—amounts to facts about the agent's organization and how robustly it copes with its environment.

John Martin Fischer (2003) agrees that there is a place for the wide reading of possibility but rejects Dennett's claim that it is the only reading that matters to "serious investigators of possibility." When we say we could have done otherwise we do not typically think that we are referring to alternate starting conditions of the universe. Rather, we believe that freedom "consists in [one's] power to add to the given past, holding the natural laws fixed" (635). The relevant sort of additive power goes beyond mere contribution. If lightning strikes a tree, igniting it and causing a forest fire, that tree does contribute to the given past—there would have been no forest fire had it not existed. Moreover, the tree's contribution depends on one of its dispositional properties (flammability), which it possesses in virtue of its physical organization. Still, the tree is not a free agent. Fischer worries that compatibilism cannot advance if it fails to at least *respect* libertarian intuitions about what freedom entails. The compatibilist may reject the standard interpretation of PAP, but in doing so they assume the burden of showing how free agents contribute to the given past in distinctly agential ways.

A related worry targets Dennett's mild realism. Dennett argues that we find the language of agency indispensible once a system achieves a certain level of complexity, but that it is a mistake to look for anything metaphysically deeper than that. But our intuitions about freedom include the idea that agents have some privileged status in the causal order of things, independent of our interpretive practices. Even if the agent is not the *ultimate source* of its behavior, it has objective properties that allow it to contribute to the progression of events in uniquely agential ways. In what follows I offer a view of agency that is compatible with Dennett's view but that privileges the role of the agent in a way that does (more) justice to our intuitions about free will.

Primitive Agency

Most discussion of agency emphasizes higher-order processes of deliberation and planning. Such discussions are valuable, but there is reason to think that a bottom-up approach to understanding action may be equally illuminating. Frankfurt (1978) argued

that a complete action theory would accommodate an active/passive distinction in animals that are incapable of deliberate action. To use his example, there is a difference between when a spider moves its leg and when its leg is moved from without. Tyler Burge (2009) argues that very basic systems such as eukaryotic cells count as primitive agents. Despite lacking capacities required for deliberative agency, there is an intuitively plausible distinction both between things those organisms do and things that happen to them, and between things organisms do and things their parts do—to use Burge's example, the *amoeba* eats but *its gullet* digests. Burge characterizes primitive agency as whole-organism functional behavior, but it is far from clear when to characterize a behavior as "whole-organism," particularly in very simple systems.

Biological interest in the whole organism as an object of study has recently surged (for a helpful summary see Nicholson 2014), but it has a long History. Kant (1987/1790) acknowledged the uniqueness of organic life in his *Critique of Judgment,* noting that their parts "are reciprocally cause and effect of their form" and that "the possibility of [the system's] parts… [must] depend on their relation to the whole" (287). Konrad Lorenz (1996/1944) defined organisms as organic *entities*, which he defined as "regulatory systems of universal, reciprocal causal connections" (137). Entities are not mere constructions of their parts because the activities of those interdependent parts are subordinated to the activities of whole entities—the parts of living systems are continually changing, and their changes are governed by the constitution and activities of the systems they comprise. Moreover, since life depends on a continuous process of endothermic assimilation and exothermic dissimilation of matter—the living system persists by breaking down its parts and rebuilding them—the whole entity displays greater invariance than its parts (85).

Recent work on self-organization offers a framework for understanding whole-system behavior. Alicia Juarerro (1999) suggests that agent-individuation amounts to identifying the proper collective variables governing the behavior of a complex system—we are "eddies of order" (145). Whole-organism behavior might be best described as those processes that correspond to changes in the order parameter values. However, it is far from obvious what ought to guide the process of order parameter selection. Indeed, from the standpoint of the complexity theorist the matter of what systems count as agents and which do not may be difficult to settle objectively—in a world of flux, boundaries may be drawn as the observer sees fit (Dennett's "mild realism" may be motivated by such considerations). Moreover, not all self-organizing systems are agents; storm clouds, tornadoes and crystal lattices cannot act. More work must be done to find agency worth having.

From Order to Value

One plausible move is to distinguish the living organism as a locus of purpose or *value*. Kant refers to self-organizing systems as "natural purposes." Lorenz distinguishes organisms from "physical gestalts" by their *finality* "in the sense of purposive survival value" (142). Jonas (1966) argues that purposiveness and value arise from the "needful freedom" of the organism, engendered by the very metabolic processes that differentiate it from its environment.[1] Jonas characterizes the phylogeny of organismic life as a series of systems that enjoy increasing freedom from their environment as they increase in complexity. The earliest form of freedom manifests in the self-organizing system's apparent violations of the Second Law of Thermodynamics—the free system first and foremost "oppos[es] in its internal autonomy the entropy rule of general causality" (5).

Jonas argues that living systems are unique among self-organizing systems in that they are essentially concerned with self-production,[2] a process through which the system distinguishes itself as *autonomous*. Definitions of biological autonomy vary widely,[3] but most characterize it as a property of far-from-equilibrium, operationally closed, dissipative self-organizing systems. This characterization can be made more concrete by examining a paradigmatic case of biological autonomy: the autopoietic system.

Maturana and Varela (1973) define living systems as autopoietic machines:[4]

1. Here Kant and Jonas break with Lorenz, who offers an evolutionary teleofunctional approach. This disagreement has no bearing on the present line of argument and will not be addressed here.

2. … living things… are unities of a manifold, not in virtue of a synthesizing perception whose object they happen to be, nor by the mere concurrence of the forces that bind their parts together, but in virtue of themselves, for the sake of themselves, and continually sustained by themselves… *This active self-integration of life alone gives substance to the term "individual."* (Jonas 1966, 79 [my emphasis]).

3. Ruiz-Mirazo and Moreno (2004) define *basic autonomy* as "the capacity of a system to manage the flow of matter and energy through it so that it can, at the same time, regulate, modify, and control: (i) internal self-constructive processes and (ii) processes of exchange with the environment." (240). Thompson (2007), citing Varela (1979), defines the autonomous system as a system whose constituent processes "(i) recursively depend on each other for the generation and their realization as a network, (ii) constitute the system as a unity in whatever domain they exist, and (iii) determine a domain of possible interactions with the environment." (44). Hooker (2011) defines autonomy as "the internally organized capacity to acquire ordered free energy from the environment and direct it to replenish dissipated cellular structures, repair or avoid damage, and to actively regulate the directing organization so as to sustain the very processes that accomplish these tasks" (35).

4. Here "machine" simply denotes systems that are defined by their organizations.

> An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces the components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as such a network. (79)

This process amounts to the self-production of the system. Crucially, the autopoietic system forms and continuously maintains a boundary, distinguishing its internal processes of the system from those of its environment. The boundary is both the product of and a necessary condition for the cell's metabolism, simultaneously limiting, contributing to and being sustained by the system's internal dynamics. Through this process of self-production and differentiation the system distinguishes itself from its environment as an autonomous unity.

The living system's "needful freedom" is due to the fact that its apparent violation of the Second Law is only apparent: it cannot remain in a far-from-equilibrium state without energy from its environment. Paradoxically, it cannot differentiate itself from its environment without continuously engaging it. Here the system's boundary serves to distinguish organization-sustaining elements from harmful elements. Furthermore, structure of the organism creates what Sørensen and Zienke (forthcoming) call an "asymmetry of normativity": depending on the structure of the system at a given time, certain features of its environment contribute to its self-maintenance—and thus are good for it—and others do not. In this way the organism's organization defines its subjective world—what the ethologist Jakob von Uexküll called its *Umwelt*.

This interaction between agent and environment evokes Merleau-Ponty's (1963) metaphor of "a keyboard which moves itself in such a way as to offer such or such of its keys to the in itself monotonous action of an external hammer" (13). Value is not an objective feature of the environment (as it might be if the structure of the musical piece were found in the environment for the passive keyboard to receive). It is constructed by the system as it navigates its environment over time. The system and its environment generate meaning through their mutual interactions at the system's boundary. Autonomous agents are not mere "eddies of order," but rather wellsprings of value.

Embedded Norms

The normative structure of even basic agency extends beyond the matter of maintaining one's organization over time. Jonas (1966) argues that the norms governing animal movement are unique within the organic world. Unlike non-motile organic systems, which either exploit the materials with which they are in direct contact or die, motile animals evolved to fit an environment wherein the materials needed for survival are spatiotemporally distant. Unlike certain plants, which can generate the materials they need to live by exploiting light energy and minerals from their soil, animals cannot manufacture the proteins, carbohydrates, fats, etc., they need on their own. So organized, animals both *must* and *can* seek out these materials in other organisms. We might say that their organizations *dynamically presuppose* this need.[5]

Jonas views this shift from basic metabolic need to what he calls *appetite* as a gradual increase in the "transcendence" of life beyond its "point-identity" (85). This "transcendence" involves a sort of dynamic presupposition along both spatial and temporal dimensions: the sensorimotor organism depends upon an energy supply that lies well beyond its boundary. Jonas notes that one consequence of this gap is a corresponding gap separating "action from its purpose" (104). Motile animals must perform "intermediate" movements, which contribute to metabolism only indirectly. These movements draw upon the organism's energy reserves, "an expenditure to be redeemed only by [its] eventual success" (ibid). Thus animal agency essentially involves gambling with one's energy reserves in hopes that the environmental payoff will have been worth the risk, for example, by funneling them into pursuit or avoidance behaviors. Stimuli are not merely "good" or "bad," but also "worth it" or "not worth it" from the perspective of a sensorimotor agent with real-time energetic needs in an uncertain environment.[6] I submit that the enaction of these "embedded" norms is the hallmark of primitive agency.

---

5. "That is, it is derived from the norm of contributing to the maintenance of the conditions for the far from equilibrium continued existence of the system… More generally, a process dynamically presupposes whatever those conditions are, *internal to the system or external to the system*, that support its being functional for the system." Beer (forthcoming) refers to the agent as prospering in "precisely those environments to whose spatiotemporal structure its autopoietic dynamics is matched" (28). It is this "match" or "fit" between agent and environment that I mean to capture with this term, applied in a different context in Bickhard (1993).

6. Here a connection can be drawn with Millikan's (1993) discussion of the embedded character of functional behavior. Millikan argues that the difference between functional behavior and other functional state changes is determined by whether the state change effects changes in the organism's environment such

How does any of this give us "freedom worth having?"

This approach motivates a biologically-oriented compatibilism that is friendly to Dennett's approach but avoids some of its shortcomings. Biologically autonomous agents are composed of matter that obeys the deterministic laws of nature. However, as a complex system the autonomous agent's causal contributions to the world are distinct from those of other objects. This is because complex systems persist by imposing constraints on their constituent parts. Atoms "caught up in the life of an organism" (Van Inwagen 1990) behave in ways they otherwise would not—their individual freedom is restricted by the system's global structure. Entrainment is common to complex systems—the vortex of water that emerges when one drains a bathtub—an *actual* "eddy of order"—is a clear case.

But as Van Inwagen notes, the living system does not "simply deposit and withdraw sequentially an invariant sum of energy" as an actual eddy might but rather "takes the energy it finds and turns it to its own purposes" (89). The world proceeds the way it does in part because the autonomous agent has the needs it does; the processes that constitute the agent's perspective and that engender its needs are doing the relevant constraining. The biologist J.Z. Young offers an apt characterization:

> The essence of a living thing is that it consists of atoms of the ordinary chemical elements... caught up into the living system and made part of it for a while. The living activity takes them up and organizes them in its characteristic way. The life of a man consists essentially in the activity he imposes upon that stuff. (1971, 86–87)

Thus the organization of a living system entrains its constituent matter into activities that serve its perspective and satisfy its needs. I have argued that in the case of the sensorimotor agent, those needs primarily involve the investment of energetic resources in adaptive sensorimotor activity. They may motivate energetic investment in pursuit behaviors at some times and avoidance at others. But it is *up to the organism* how its resources are invested—other organisms with distinct structures might have behaved quite differently in the same circumstances.

We may now revisit Dennett's claims about what an organism could have done. The autonomous agent's organization at any given time determines the range of

---

that the organism gets a return on its investment. This is why, for example, the clam's slowing its activity in cold water does not count as behavior but the spider's pursuit behavior does: only the latter involves an energetic *investment*, rather than a mere expenditure.

environmental features to which it will be receptive (effectively determining the range of possible past timelines that will matter for its action). Its organization also determines the trajectory of future causal events: the agent is successful when it is able to steer that trajectory in ways that satisfy its needs; it fails when it cannot. We can judge the agent's capacities by appeal to its robustness across possible timelines: the range of alternate causal trajectories that it can steer in its favor. To say that the rabbit could have avoided the hawk is to suggest that there is a range of possible deterministic unfoldings of the universe that happen to include that rabbit at that time, and that in some of those unfoldings the rabbit's organization successfully channels enough energy into the task of avoidance.

None of this denies our intuition that freedom consists in a distinctly agential power to add to a given past. The processes that channel matter and energy through the organism operate as they do *because* of the organism's perspective and needs. The agent may not be the *ultimate* source of its behavior—this is, perhaps, too much to ask—but it does matter in a distinctly agential way. I submit that this is a form of free will worth having.

References

Beer, Randall. 2004. "Autopoiesis and Cognition in the Game of Life." *Artificial Life* 10 (3): 309–326.

Bickhard, Mark. 2004. "The Dynamic Emergence of Representation." In *Representation in Mind: New Approaches to Mental Representation*, edited by H.Clapin, P. Staines and P. Slezak, 71–90. Oxford, UK: Elsevier Inc.

Burge, Tyler. 2009. "Primitive Agency and Natural Norms." *Philosophy and Phenomenological Research* 79 (2): 251–278.

Dennett, Daniel. 2003. *Freedom Evolves*. New York: Viking.

Di Paolo, Ezequiel. 2005. "Autopoiesis, Adaptivity, Teleology, Agency." *Phenomenology and the Cognitive Sciences* 4 (4): 429–452.

Fischer, John. 2003. "Review of Freedom Evolves by Daniel C. Dennett." *Journal of Philosophy* 100 (12): 632–637.

Frankfurt, Harry. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66 (23): 829–39.

Frankfurt, Harry. 1978. "The Problem of Action." *American Philosophical Quarterly* 15 (2): 157–162.

Hooker, Cliff. 2011. *Handbook of the Philosophy of Science. Volume 10: Philosophy of Complex Systems.* Waltham, MA: Elsevier B.V.

Inwagen, Peter. 1990. *Material Beings*. Ithaca, N.Y.: Cornell University Press.

Jonas, Hans. 1966. *The Phenomenon of Life: Toward a Philosophical Biology*. New York, NY: Harper & Row Publishers, Inc.

Juarrero, Alicia. 1999. *Dynamics in Action: Intentional Behavior as a Complex System*. Cambridge: MIT Press.

Kant, Immanuel. 1987. *Critique of Judgment.* Trans. W.S. Pluhar. Indianapolis, IN: Hackett Publishing Company.

Lorenz, Konrad. 1996. *The Natural Science of the Human Species: An Introduction to Comparative Behavioral Research. The "Russian Manuscript" (1944–1948)*. Cambridge: MIT Press.

Maturana, Humberto, and Francisco Varela. 1980. *Autopoiesis and Cognition: the Realization of the Living.* Dordrecht: D. Reidel Pub. Co.

Merleau-Ponty, Maurice. 1963. *The Structure of Behavior*. Boston: Beacon Press.

Nicholson, Daniel. 2014. "The Return of the Organism as a Fundamental Explanatory Concept in Biology." *Philosophy Compass* 9 (5): 347–359.

Ruiz-Mirazo, Kepa, and Alvaro Moreno. 2004. "Basic Autonomy as a Fundamental Step in the Synthesis of Life." *Artificial Life* 10 (3): 235–259.

Sørensen, Mikkel, and Tom Zienke. Forthcoming. "Agents Without Agency?" *Cognitive Semiotics*.

Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Belknap Press.

Varela, Francisco, Evan Thompson, and Eleanor Rosch. 1992. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Young, John. 1971. *An Introduction to the Study of Man*. Oxford: Oxford University Press.