Guilty Robots, Happy Dogs: The Question of Alien Minds

Author: Dennett, Daniel C.

Source: BioScience, 59(8) : 707-709

Published By: American Institute of Biological Sciences

URL: https://doi.org/10.1525/bio.2009.59.8.14

learn to appreciate some of the advances and controversies in evolutionary developmental biology while reading Greg Wray. Then again, there is no essay on the role of phenotypic plasticity in evolution, a topic that has acquired central status during the past two decades; after perusing *Evolution* a reader might be excused for not appreciating the entire field of evolutionary genomics, or for being ignorant of ongoing discussions on crucial new concepts like evolvability. Even attempts to move beyond strict biology with entries on evolution and society, evolution and religion, and the above-mentioned essay on antievolutionism barely scratch the surface—why is there no discussion of evolutionary psychology, as controversial and somewhat dubious as the field is?

While some of these lacunae could have been avoided during the planning stages of the volume, I think the underlying problem is that encyclopedic efforts are a thing of the past, certainly when it comes to the paper variety of encyclopedia. In this bold new era of ubiquitous and increasingly cheap laptop computers, 24/7 Internet access, e-readers, smart phones, and so on, I simply do not see many people willing to lug around a thousand pages of what is going to be a necessarily incomplete and increasingly unrepresentative reference source like *Evolution.* Publishers, editors, and authors would be much better off embracing the anarchy and flexibility of the Web to develop decentralized and more focused projects, such as the excellent *Complete Works of Charles Darwin* online (*http://darwin-online.org.uk/*).

Even encyclopedias are taking a decidedly different form these days, and if one does not like the proletarian *Wikipedia,* excellent models of scholarly efforts are out there, such as the *Stanford Encyclopedia of Philosophy* (*http://plato.stanford.edu/*). These take seriously the idea of organic, grassroots growth arising from the efforts of a dedicated community, based on what the community itself sees as worth writing about, as opposed to the centralized planning typical of the standard model. Indeed, let me suggest to Ruse and Travis, both of whom I know and highly respect, that they go back to Harvard Press and propose to use the current version of their book as the seed for a community-wide, online, open-ended effort. Of course, it would also be nice if it were open access, but that's another story.

MASSIMO PIGLIUCCI
Massimo Pigliucci (massimo @platofootnote.org) is a professor in the Department of Philosophy at the City University of New York, Lehman College.

References cited

Browne J, Ekman P, Kauffman S, May R, Pigliucci M, and Raison C. 2009. Darwin's Descendants. New York Academy of Sciences. (7 August 2009; *www.nyas.org/Publications/Detail.aspx?cid=56e35057-c0ad-4b83-8c02-bc38082419dc*)

Coyne JA. 2009. Why Evolution Is True. Viking.

Diderot D, d'Alembert JR. 1751–1777. Encyclopédie. (7 August 2009; *http://quod.lib.umich.edu/d/did/*)

Müller GB. 2007. Evo-devo: Extending the evolutionary synthesis. Nature Reviews Genetics 8: 943–949.

Norenzayan A, Shariff AF. 2008. The origin and evolution of religious prosociality. Science 322: 58–62.

Pigliucci M. 2007. Do we need an extended evolutionary synthesis? Evolution 61: 2743–2749.

WHAT IS IT LIKE TO BE A ROBOT?

Guilty Robots, Happy Dogs: The Question of Alien Minds. David McFarland. Oxford University Press, 2009. 256 pp., illus. $15.95 (ISBN 9780199219308 paper).

Any scientist who wants to investigate minds—our minds, animal minds, alien minds—will soon discover that there is no way to proceed without venturing into the playgrounds and battlefields of the philosophers. You can either stumble into this investigation and thrash about with a big scientific stick, thwacking yourself about as often as your opponents, or you can enter cautiously, methodically, trying to figure out the terrain using what you already know to interpret what you find. Fortunately, David McFarland has chosen the second option in *Guilty Robots, Happy Dogs: The Question of Alien Minds,* and there is much food for thought here for both scientists and philosophers.

doi:10.1525/bio.2009.59.8.14

It is written in the spirit of Valentino Braitenberg's brilliant little book *Vehicles* (1984), a series of thought experiments that led readers from robotic vehicles even simpler than bacteria to ever-more sophisticated and versatile agents capable of tracking food, avoiding harm, comparing situations, and remembering things. McFarland starts his project a little higher on the ladder of sophistication, with a robot designed to serve as a night watchman of sorts, identifying interlopers, calling for help when needed, and, most important, preserving its energy supply for another day, budgeting its activities to stay alive at all costs. This basic robot is then enhanced in various ways, in a design process whose ultimate goal is a robot that can be held accountable and to whom things matter—a robot with subjectivity and values.

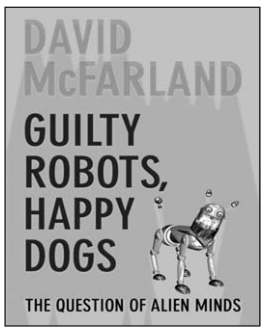
How do nonhuman animals compare with such robots? Animal minds (including our own) are the real quarry here, and McFarland uses the parallels and differences between clearly imagined robots and various well-studied animals to illuminate the issues in a host of research controversies currently raging in psychology and ethology. This has been his larger strategy for many years, and this book gives us a summary of the lessons he has gleaned from this interdisciplinary exploration.

One message driven home most effectively, in my opinion, is that it is entirely appropriate to consider natural selection to be a (mindless, purposeless) designer, and to compare the designs churned up by eons of natural selection on a par with designs generated top-down by would-be intelligent designers—engineers and roboticists. Sometimes the perspective is particularly bracing, as when McFarland insists on situating his imagined robots in a market economy so he can note

that nobody would buy such a robot—it wouldn't pay for itself. Animals, similarly, are amazingly thrifty because they have to be; they have superb layers of self-protection and repertoires of self-advancing behaviors, but not a smidgen more than can pay for itself in the long run. This often brings out the rationale for animal (or robot) features that would otherwise be lost in the shadows. It also obliges McFarland to commit to a "behaviorist" approach—not the ideological straitjacket of the Skinnerians but the behaviorism expressed by Turing in 1937, when he noted about the human computers of his day: "The behavior of the computer at any moment is determined by the symbols which he is observing, and his 'state of mind' at that moment" (p. 241). Handsome is as handsome does, a motto enshrined in the rationale for the Turing test, and the only way a *science* of mind can proceed.

But how much can one glean from inner behavior (the machinery of the mind, in effect) by observing the competencies of outer behavior? Do animals, for instance, really have the beliefs that their behavior seems to indicate they do? Here McFarland avails himself of a slightly unorthodox but useful interpretation of philosophy's terms of art, realism and functionalism. Realism would not impute a belief to the organism unless it was "in principle identifiable outside the role that it plays in the system" (p. 69), whereas functionalism (such as my intentional-stance view) is more relaxed, willing to impute beliefs that are only implicit in the design and functioning of the larger system. For realists, a belief is an explicit representation, "not simply part of a procedure. If a representation is to be explicit, then there has to be a physically identifiable bearer of the information (the token) and, additionally, something, most likely someone, who can be identified as the user of the information" (p. 77).

Human beings have beliefs aplenty, obviously, because they have lots of explicit knowledge that they can readily express. Do dogs or robots have explicit beliefs? Do they need them? McFarland shows how robots can exhibit behaviors similar to animals' behaviors without explicit representation, and he proposes to define cognitive processes as those that require "a certain kind of mechanism—one that requires manipulation of explicit representations" (p. 87). This sets the bar high and departs from standard usage, but perhaps it is best to follow his lead. Note that with this definition, it isn't clear that our hand-eye coordination or even our ability to find our way home counts as a cognitive process (unless we use a map or an explicit mental map).



McFarland also proposes a demanding definition of subjective experience: "The agent is the recipient of experience, and knows it" (p. 94). Using this definition, the behavior of turning to a more painless posture while asleep would not count as demonstrating subjective experience of pain, and it follows that much animal behavior (think of fish, for instance) is not clear evidence that animals have subjective experience, no matter how frantically they squirm. McFarland does not infer that animals don't have subjective experience or explicit beliefs. He just points out that given these well-motivated definitions, we cannot yet tell.

Indeed, that is the larger conclusion that McFarland draws again and again—the evidence is not yet in, not even about Border, his dog. He looks sympathetically at important experiments and observations, of dogs "teaching" their pups, of animals making sophisticated choices (are they explicitly maximizing their expected pleasure?). In each case he finds that a functionalist interpretation of the behavior seems to suffice: "Certainly we can say that the teacher behaves as if it wants, hopes, or desires the pupil to behave in a certain way," he says, but he also goes on to note that the teacher could have a "strong theory of mind" about the pupil and be wrong (p. 105). The comparison with robots is always astringent here, and McFarland puts our built-in skepticism about robot minds to good use in reining in our romanticism about our furry friends.

McFarland proposes a contrast between two views of what is going on inside: the hedonic model and the automaton model. According to the hedonic, "the feelings of pleasure and displeasure that arise from various parts of the body in situations of motivational compromise are combined in some way, and behavioral adjustments are made so as to maximize pleasure and minimize displeasure." By contrast, in the automaton, "the behavioral and physiological adjustments are automatic, and...the system is attuned to produce the best compromise among the competing demands" (p. 123). He says, "The fundamental difference between the two views is that in the automaton view the quantity maximized is implicit, while in the hedonic view it is explicit" (p. 123). But are these views as distinct as they first appear? When he turns to Michel Cabanac's experiments with people being paid to endure discomfort, and paying for sandwiches of different tastiness (by their own taste), he can rely, for once, on what subjects say about their decisions. As he goes on to note, however, a subliminal process can take the place of a "conscious motive," apparently, and thus "it is not clear that Michel Cabanac is correct in assuming that trade-offs involving money necessarily involve a conscious mental component" (p. 128).

As we near the summit, we consider robot designers who want their robot to be "accountable for its behavior." For this, it needs its own values, not just its designers' values. It can learn to adjust its values, but this learning must depend on some prior "immutable" values it was born with, you might say. Here I think McFarland misses a possibility: It

might be unwise to design a robot that could eventually undo even its initial "immutable" values and take on a new *summum bonum,* but this is not an engineering impossibility (Suber 2001). Perhaps the only way to make an accountable robot that could deserve punishment for its misdeeds and rewards for its heroics would be to give it the dangerous capacity to renounce the values we installed in it at birth.

McFarland has done his homework well; he offers a patient, sympathetic, and largely accurate discussion of philosophers' best relevant work, plunging into the darkest thickets of controversy over supervenience, eliminativism, symbol grounding, higher-order thought theories, and the like. Some of his readings will jar the authors he discusses, who will think that they have been misunderstood to hold positions that had never occurred to them, but they will never find him sniping in standard philosophical fashion; if he misreads them, it is because his effort to find a constructive reading was too charitable by half. Philosophers are not always trying to do as much as scientists imagine.

DANIEL C. DENNETT
Daniel C. Dennett (daniel.dennett@tufts.edu) is the Austin B. Fletcher Professor of Philosopy and codirector of the Center for Cognitive Studies at Tufts University in Medford, Massachusetts.

References cited

Braitenberg V. 1984. Vehicles: Experiments in Synthetic Psychology. MIT Press.

Suber P. 2001. Saving Machines from Themselves: The Ethics of Deep Self-Modification. (22 July 2009: *www.earlham.edu/~peters/writing/selfmod.htm*)

Turing A. 1937. On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society 42: 230–265. Erratum in Proceedings of the London Mathematical Society 43: 544–546 (1938). doi:10.1112/plms/s2-43.6.544

LIFE AT THE ENDS OF THE EARTH

The Biology of Polar Regions. 2nd ed. D. N. Thomas, G. E. Fogg, P. Convey, C. H. Fritsen, J.-M. Gili, R. Gradinger, J. Laybourn-Parry, K. Reid, and D. W. H. Walton. Oxford University Press, 2008. 416 pp., illus. $60.00 (ISBN 9780199298136 paper).

As a scientist who has spent 25 years conducting research in polar regions, I was immediately drawn to *The Biology of Polar Regions* for several reasons: (a) The poles, particularly Antarctica, represent one of the last frontiers of exploration on our planet; (b) polar environments are highly sensitive barometers of climate change and can affect the entire Earth system as they respond to changing climate; and (c) this work continues the vision and style of the late G. E. Fogg, who was not only an eminent scientist but also a visionary able to view Earth in a completely holistic way. The first version of the book, published in 1998, was written by Fogg alone; for this second edition, it took eight authoritative authors—with expertise spanning topics such as marine biology, biological oceanography, sea ice, soils, limnology, climate change, and Antarctic conservation and policy—to update Fogg's original version. Kudos to the present authors for maintaining Fogg's original chapters 1 and 2 and the concluding chapter largely unchanged, paying tribute to his inquisitive pursuit of the nature of science and masterful synthesis of information across many disciplines.

Fogg's first two chapters describe the basic physical and biological constraints on life in polar regions. The subsequent nine chapters are updated substantially to reflect the many new discoveries made over the past decade. I found the chapters on sea ice, marine benthos, and human impacts (particularly the review of polar politics) to be exceptional. Any student of the polar sciences must read these chapters. Difficult and often controversial topics such as photophysiology (chapter 2) and turbulence (chapter 3) provide the reader with a nice review of the processes before delving into their roles in polar environments. Each chapter typically ends with summary or conclusion sections, which could work well, but, unfortunately, several of these sections are not well developed. For example, the "wider perspectives" section of chapter 4 ("Glacial Habitats in Polar Regions") hardly goes beyond the surface on the role of polar environments as analogs for life on other icy worlds. Only two references in this section were published after 2005, which does little justice to what we have learned about the icy systems of Mars, Europa, and Enceledus over the past five years. I think that many aspects of this chapter, and the book in general, are directly relevant to astrobiology and worthy of better coverage.

doi:10.1525/bio.2009.59.8.15

This edition of The Biology of Polar Regions *packs a plethora of information. The authors' detailed comparisons of Arctic and Antarctic habitats generate a breadth of coverage that few books on polar environments offer.*

I was also left wondering to what degree the data presented in many of the chapters have changed over the past five years as the result of climate change. This is particularly relevant to the descriptions of species within Arctic terrestrial and marine systems, which are clearly in a state of transition. For example, are the depth profiles, species lists, and food-chain depictions in chapter 5 ("Inland Waters in Polar Regions") and chapter 6 ("Open Oceans in Polar Regions") representative of the present situation? Current literature indicates that these relationships have changed or are in the process of changing rapidly. Although the authors do devote chapter 10 to describing the influence of climate change on many levels of polar ecosystems, and even include a section on