4

The Baldwin Effect: A Crane, Not a Skyhook

Daniel Dennett

1 Introduction

In 1991, I included a brief discussion of the Baldwin effect in my account of the evolution of human consciousness, thinking I was introducing to non-specialist readers a little-appreciated, but no longer controversial, wrinkle in orthodox neo-Darwinism. I had thought that Hinton and Nowlan (1987) and Maynard Smith (1987) had shown clearly and succinctly how and why it worked, and had restored the neglected concept to grace. Here is how I put it then:

> If we give individuals a variable chance to hit upon (and then "recognize" and "cling to") the Good Trick in the course of their lifetimes, the near-invisible needle in the haystack . . . becomes the summit of a quite visible hill that natural selection can climb. . . . Over generations, the competition becomes stiffer: eventually, unless you are born with (or very nearly with) the Good Trick, you are not close enough to compete. If it weren't for the plasticity, however, the effect wouldn't be there, for "a miss is as good as a mile" *unless* you get to keep trying variations until you get it right.
>
> Thanks to the Baldwin effect, species can be said to pretest the efficacy of particular different designs by phenotypic (individual) exploration of the space of nearby possibilities. If a particularly winning setting is thereby discovered, this discovery will *create* a new selection pressure: organisms that are closer in the adaptive landscape to that discovery will have a clear advantage over those more distant. This means that species with plasticity will *tend* to evolve faster (and more "clearsightedly") than those without it. (Dennett 1991: 186)

I now discover that there are still "Baldwin skeptics" (Downes, Godfrey-Smith, chaps. 2 and 3, respectively, this volume) who do not so much doubt the possibility of the Baldwin effect as doubt its importance as what I have called a *crane:* "a subprocess or special feature of a design process that can

be demonstrated to permit the local speeding up of the basic, slow process of natural selection, *and* that can be demonstrated to be itself the predictable (or retrospectively explicable) product of the basic process" (Dennett 1995: 76). This is what I claimed:

> It shows how the "blind" process of the basic phenomenon of natural selection can be abetted by a limited amount of "look-ahead" in the activities of individual organisms, which create fitness differences that natural selection can then act upon. (Ibid.: 80)

So the question is: Does the Baldwin effect really occur, and if it does, how much "local lifting" can it do? Can it make a significant difference in the trajectories through Design Space that species have traversed? Part of the problem is a relatively unimportant wrangle about naming and boundaries: shouldn't the Baldwin effect be named after somebody else, and is the Baldwin effect just a special case of Waddington's (or somebody else's) "genetic assimilation," and if so, how special is it, if at all? The rest of the problem concerns the soundness or realism of the (idealized, oversimplified) models and explanations, and the more directly empirical question of whether in fact there are demonstrated cases of Baldwin effects in nature. As so often happens in evolutionary controversies, the empirical questions of most interest have to be traded in for something more practical if we want to test them experimentally. We'd love to know about the role of language-learning abilities in *H. sapiens* and its ancestors, but we have to settle for the more modest learning (if that's not too strong a term) abilities of something like *E. coli* or *D. melanogaster* if we want to look at hundreds or thousands of generations. I expect that experiments will eventually shed light on this, if they haven't already done so (I confess that I am not up to date on the literature), but in the meantime, several commentators have expressed objections to the more strictly theoretical discussions that I want to answer.

2 Peering Through the Fog of Battle

These theoretical objections are valuable because they expose a vein of misdirection that continues to haunt evolutionary thinking a century and more after Baldwin mounted his campaign to claim the discovery as his own. That campaign finally succeeded more than a half century after it started,

when George Gaylord Simpson (1953) dubbed it the Baldwin effect, but it was presumably a pyrrhic victory, since, as Downes (chap. 2, this volume) notes, Simpson coined the term "in order to argue it did not exist as an independent factor from natural selection." In fact, Simpson does not so much argue for this deflationary conclusion as simply express his (expert) opinion. He demonstrates that the Baldwin effect is possible, calling it "an interesting, but, I would judge, relatively minor outcome of the theory" (Belew and Mitchell 1996: 107), and goes on to remark that Waddington's genetic assimilation represents "a broader principle of which the Baldwin effect may be considered a special case" (ibid.). The last sentence of Simpson's short paper reveals that his main target in writing the paper has been to dump cold water on a misguided enthusiasm: "It does not, however, seem to require any modification of the opinion that the *directive force* in adaptation, in the Baldwin effect or in any other particular way, is natural selection" (108).

From the outset, Baldwin had advertised the effect as an instance of what Downes calls "mind-directed" evolution, and it was gullibility about this prospect that Simpson was trying to squelch; but Baldwin's own account (see especially his discussion of the possible mechanisms of "selection" in what he calls "neurogenetic" modifications, p. 61 in Belew and Mitchell 1996) shows that he was always alert to the requirement that his commitment to Darwinism obliged him to postulate a crane, not a *skyhook* ("a 'mind-first' force or power or process, an exception to the principle that all design, and apparent design, is ultimately the result of mindless, motiveless mechanicity," Dennett 1995: 76). Baldwin's own presentation of the issue was thus Janus-faced: he seemed to promise a skyhook to the eager buyers but was careful to deliver an orthodox crane, a fact underappreciated by some who took up the cause. Godfrey-Smith closes his paper by noting that Simpson, in his eagerness to preserve orthodoxy from contamination, also seems to have overshot his target somewhat in the opposite, compensatory direction: "the dichotomy is either false, or else Simpson is wrong that a causal connection of the kind he describes is not compatible with mainstream Darwinism."

The same tug of war distorts the rhetorical setting of Stephen Downes's latter day skepticism: if the Baldwin effect is "just" a special case of genetic assimilation, then it is not a "new evolutionary mechanism" (Downes, this

volume) and hence no big deal; if, on the contrary, it is supposed to be an *alternative* to natural selection, an independent source of "mind-directed evolution," then it is (deservedly) suspect and contentious. My claim continues to be that the Baldwin effect is not at all an alternative to natural selection, but it is nonetheless an important extrapolation from, or extension of, orthodox theory that potentially can explain the origins of many of the most challenging adaptations. And so when Downes goes on to say (this volume) "It may be safer to say that the non-Waddington style concept of genetic assimilation may account for a subclass of phenomena labeled as Baldwin effects," he is not disagreeing with me, with Dawkins, with Deacon, or with Simpson. That is the bland position that we, too, take.

Downes (this volume) tries to manufacture a conflict between my claim that this provides a limited "look-ahead" and my agreement with Williams (and orthodoxy) that there is no foresight in natural selection, but this charge of inconsistency readily collapses: design explorations by phenotypic trial and error are just as mechanical and nonmiraculous as explorations by genetic natural selection; they just occur more swiftly and at less cost, and once design improvements are thereby discovered, genetic assimilation can incorporate them gradually into the genome. I wonder: does Downes think that the existence of genetic engineers is a problem for orthodox neo-Darwinism? These people now second-guess evolution on a broad front (with mixed results, of course, but surely it is better than a coin toss, better than random). How are these people capable of any foresight when they do this? Are they themselves gods, not products of natural selection? Of course not. It is obvious, I would have thought, that this look-ahead is itself the product of conscious and deliberate human reasoning and analysis, which is itself a product of earlier evolutionary processes. Our capacity to look ahead is as uncontroversially real as our capacity to breathe and metabolize. It had to evolve.

Moreover, contrary to Downes's discussion, my point has always been to stress that learning is just a particular case (not in any other way special) of ontogenetic adaptation. The continuity between learning and other purported varieties of self-redesign is taken as given in the circles in which I converse; learning is adaptive, functional change of one's cognitive (or control) mechanisms, as contrasted with one's digestive mechanisms, reproductive mechanisms, and so on. Hence no part of my purpose was to

propose any sort of threshold distinguishing learning as a distinct phenomenon. (Godfrey-Smith, chap. 3, this volume, follows the same familiar policy when he says he is "going to use 'learning' as a short-hand for all facultative mechanisms for acquiring traits.") Wherein lies the importance of the Baldwin effect, then, if it is "just" business-as-usual ontogenetic adaptive plasticity leading the way to genetic adaptation?

It is not that the Baldwin effect accounts for otherwise inexplicable differences in tempo in evolution, but that it accounts, as Maynard Smith so crisply shows, for the evolution, in sexually reproducing species, of traits that theory would otherwise declare to be all but unevolvable—those needles in a haystack that would otherwise be invisible to natural selection. The importance of this issue does not loom large for either Godfrey-Smith or Downes. I don't know why. Perhaps it is because they, like many others, have been taught at least to feign discomfort when adopting the adaptationist perspective, or perhaps because they have not encountered much of the bizarre skepticism regarding the evolution of language (and "language acquisition devices") that has haunted the corridors of linguistics and philosophy of mind over the years. Putting the best interpretation on this skepticism (that is to say, ignoring the sometimes highly tempting diagnosis of closet Creationism), it amounts to a general conviction that something as specialized as the imagined "language acquisition device" is just such a needle in the haystack, something that could not evolve gradually but would have to be an almost miraculous saltation, a cosmic accident of good luck—what a Creationist would call a gift from God. Nonsense, say we Baldwin effect supporters. A practice that is both learnable (with effort) and highly advantageous once learned *can* become more and more easily learned, can move gradually into the status of not needing to be learned at all. It is instructive to note the parallel between this battlefield and the ground on which Waddington mounted his campaign for genetic assimilation: how *could* the embryonic callosities on ostrich legs (and human soles) be explained by orthodox Darwinism without appeal to Lamarckian mechanisms? In both cases, the initial, superficially plausible incredulity or skepticism must give way to an appreciation that evolution has a few more tricks up its sleeve than heretofore imagined; *there are* paths of (non-Lamarckian) orthodoxy leading from adaptative phenotypic adjustments to inherited genetic arrangements.

3 Trade-offs between Learning and "Instinct"

As Godfrey-Smith and others have noted, the purported outcome of Baldwin effects is *reduced* phenotypic plasticity (for the trait in question), so the Baldwin effect cannot be trundled out to explain the evolution of learning. Nevertheless, there is need for an account of the relationship between selection pressures in favor of enhanced learning abilities and selection pressures in favor of driving a new trick into the genome. Consider the generalized case in which the Baldwin effect is supposed to operate.

When a new Good Trick is discovered (by some member or members of a population), any genetic variation in the population that makes the learning swifter or more probable should have a fitness advantage, other things being equal. Different sorts of variation may happen to exist simultaneously in the gene pool, operating in two quite different ways:

(A) giving a leg up: starting the individual off in a state closer *in learning space* to the mature practice, so there's simply less to learn (this is the Hinton and Nowlan variation); and

(B) putting more spring in the legs: enhancing the learning capacity itself, so that the "lifting" distance is more swiftly and surely covered (this is variation in learning ability or adaptability).

Of these two "opposite" paths—one heading toward creating a new "instinct" and the other heading toward creating greater "general intelligence"—which will be favored? Presumably the incidental costs and benefits in each case will tip the scales one way or the other, and this is plausibly a highly sensitive variable. If the Good Trick has a fairly stereotypic set of releasers and conditions in the prevailing environment, and there are few *other* Good Tricks in the neighborhood it behooves one to learn, then probably the path to adding a new instinct is favored. In a more volatile environment, the costs of working harder to get the Good Trick may have enough incidental side payoffs to favor maintaining, and enhancing, the learning machinery instead. In some circumstances a species would be wise/lucky to "pay" for this increased learning *speed* by moving the neonate *farther* away, in learning space, from the Good Trick. This tidy picture is no doubt complicated in reality by dozens of other effects that might swamp this underlying consideration: perhaps a particular anatomical de-

tail in some brains makes certain sorts of learning (or instinct) particularly expensive; perhaps there's an interaction with metabolism or growth rate or who knows what else. In any event, this saddle in Design Space must have often confronted species, for we see a host of instances in which what is fixed and instinctual in one species is variable but learnable in another.

Godfrey-Smith makes the point that Hinton and Nowlan's model has a particularly strong idealization in it, which he calls the Waddington requirement: "the genetic path leading through better and better learners is *also* a path leading to a well-adapted nonlearner" (this volume). The learning space is simply declared to be superimposed on the genetic space, so that there is a one-to-one mapping of mutations onto lessons-learned. This simplifies the phenomenon, since it treats the paths of learning and genetic transition as common and interchangeable; an organism can be n bits away from the Good Trick, a distance that can be traversed by any combination of learning and mutation. "As one traverses genetic space through genotypes that are more and more effective at learning a given behavior, one is also moving closer to genotypes that tend to produce the behavior without need for learning" (this volume). This may seem to be a huge and deeply unwarranted oversimplifying assumption, since it ignores what might seem to be a very real, even likely possibility: in order to "traverse genetic space" in the direction of more and more effective learning, you might have to leapfrog around in actual genetic distances. There is no guarantee, it might seem, that genotypes that are neighbors in genetic space are also similar in learning space. But in fact, Hinton and Nowlan's simplification is, so far as I can see, innocent, since it generalizes over the more realistic cases. To see this, suppose that there are, in some instance, three genetically distinct peaks in the adaptive landscape (rather distant from each other in genetic space)—three "different ways" to have an *instinct* for a specific behavior that is, at the outset, a *learned* behavior of some value. If there is selection pressure for learning the Good Trick *one way or another,* there will be simultaneous selection pressure felt on the slopes of all three peaks. The fact that there is no gradual upward path connecting all regions of genetic space to a single summit (Hinton and Nowlan's idealization) means only that there is no guarantee that there aren't suboptimal dead-end paths that must be traversed and then eventually discarded (unless a dimorphism or multimorphism happens to be stable). But we can be sure that there are *local*

gradients in favor of heightened ease-of-learning because, if we imagine holding the learning mechanism constant (whatever it is), any small change in genetic space that changes the starting point in a way that happens to shorten the distance in learning space must be simultaneously a (small) step in the right direction both genetically and phenotypically. If there are *no* such changes, then, of course, there will be no genetic assimilation, but it does not seem extravagant to suppose that there will often be a winding upward path of small steps in genetic space that have the effect of shortening the distance in learning space one way or another. Where the genetic path stops, leaving the rest of the redesign trajectory to individual learning, is then a matter that can vary indefinitely.

A thought experiment can highlight the point at issue.[1] Imagine an obsessive Skinnerian who has joined forces with an evolutionary biologist in order to create a subspecies of African gray parrots with the *innate instinct* for uttering, without any special training, without so much as hearing an exemplar, "Boo Chomsky!" There is plenty of genetic diversity among African gray parrots, and no doubt some of it is in the desired direction of a bird who would be born wanting to utter "Boo Chomsky!" at its earliest opportunity, but how on earth could the birds with these alleles be identified? How could evolution, even with a helping hand from our artificial selector, find the leverage to steer a lineage in this direction? This is where the Baldwin effect comes to the rescue. We know that African gray parrots, like mynah birds and a few other species, are particularly trainable, and even self-trainable, aural mimics, so there is no question that any Skinnerian who set himself the task of creating a flock of parrots who all said "Boo Chomsky!" would soon be able, by the "shaping" method of operant conditioning, to create an avian chorus of adult birds with just this talent.[2] Having done so, he could begin, with the help of the evolutionary biologist, to raise the bar: only those birds who were particularly trainable, the champion learners of this phrase, would be allowed to reproduce in the next generation. There would be, as we have just seen, no guarantee that this would head in the right direction. It might be that there was no way to select for the talent for saying "Boo Chomsky!" (path A) that wasn't just selecting for virtuoso trainability in general (path B). But if there were any variability down path A, if (in other words) some of the accessible genotypes were not better learners in general but just more likely to learn to say "Boo Chom-

sky!" (and perhaps a few other phrases unimagined and untested by the selectors) easier than their rivals, they would be identified by this new selection pressure *that comes into existence only when the learned adult competence is discernible in the population.* They could then be selected for exactly this proclivity, which could, in due course, go all the way to a hair-trigger utterance of "Boo Chomsky!" in need of no training at all.

Is this impossible? Who knows? It may just be, perhaps, that in the Vast catalogue of possible but as yet unrealized African gray parrot genomes, not a single one yields a bird (under normal developmental conditions) that squawks "Boo Chomsky" as soon as it squawks anything. But if there is even one such genome, the beauty of the Baldwin effect is that it shows that there is no theoretical reason to rule out the otherwise astonishing feat of finding such a genetic needle in the haystack. Anyone who was tempted to assert that there is simply no way for natural selection to produce an African gray parrot whose "instinctual call" is "Boo Chomsky!" would be ignoring a path of orthodoxy, an unappreciated mechanism, but not a new or revolutionary one. This thought experiment of mine will have to suffice for the time being as my response to Downes's challenge for me to produce an actual example of "phenomena that do not succumb to standard evolutionary explanations" (this volume). If any funding agency wants me to turn my thought experiment into a real experiment, I'm sure I can still find a behaviorist or two who would enjoy coming out of retirement and setting this in motion. It will take quite a few years, and be quite expensive, but it might be worth it, if it would convince the diehard skeptics. The important point is that the chances of selectively breeding such a bird depend on the bird's having enough of a "mind" to be *trainable* to utter the sounds. There might also be an unrealized genome in the Vast catalogue of possible blue jay genomes that would yield the same vocalization instinct, but it is hard even to imagine a feasible path that could take us there, since there would be no adaptive slope to guide our search.

I take it that I am so far just elaborating the standard presumption about how the Baldwin effect works, not breaking new ground. But I have learned to be cautious about this. Is Godfrey-Smith right that Deacon, in his discussion of the Baldwin effect, has added a new mechanism? Now it is my turn to play skeptic; I think Deacon is right about the heightened selection pressure brought about by the prevalence of the Good Trick in the population,

but I thought that was implicit all along in the earlier discussions—in Hinton and Nowlan, in Maynard Smith, and in my own remark (quoted above): "Over generations, the competition becomes stiffer: eventually, unless you are born with (or very nearly with) the Good Trick, you are not close enough to compete." Perhaps Deacon spelled it out better, or perhaps he has indeed proposed a new mechanism, but if so, I don't yet see what it is.

Finally, I was puzzled by Downes's dismissive suggestion regarding the evolution of lactose tolerance among people engaged in herding: "The relevant details are probably best spelled out in the molecular biology of enzyme production." But surely the difference between human subpopulations that plays the major role in explaining the observed differences in "enzyme production" is the large difference in diet, which is itself explained by a food-gathering practice that is learned, not genetically transmitted. This is not yet the Baldwin effect, but it is definitely an instance, contrary to what Downes says, of "mind directing evolution" in the bland but important sense of a learned, culturally transmitted practice having dramatic genetic consequences. Moreover, the case is pretty strong for an *indirect* Baldwin effect arising from such practices in another species involved in them. A border collie puppy hardly has to be taught to herd sheep—its instinctual skills are merely honed by training (unlike, say, the children of Basque shepherds who are not similarly genetically equipped with herding instincts!). What drove the evolution of herding instincts in border collies? The learned human Good Trick of animal husbandry. Dogs that could more readily learn to herd had a huge selective advantage, but only because of their interactions with their foresighted, looking-ahead "masters." In this case, it was the "mind-directed" activities of another species that created the gradients up which first unconscious, and later, artificial selection could drive the genomes of those wolf-kins.[3]

Acknowledgments

I am indebted to Mateo Mameli and Stephen Downes for valuable discussion on an earlier draft of this paper.

Notes

1. Since devising this case, I have been delighted to discover that a strikingly similar example with a slightly different emphasis was invented over a century ago by

Spalding (1873) quoted in Avital and Jablonca (2000): 321, a striking case of convergent evolution. Spalding's version has some variations of its own that are particularly amusing, since he supposes that sexual selection maintains the instinct long after the death of the Crusoe, the behaviorist trainer in his version. This is genuine possibility, I think.

2. A better technique that classical operant conditioning would be the ingenious imitation method used by Pepperberg (2000) in training her virtuoso vocalizer, Alex, who watches a rival being trained and competes for her attention and approval. But for the sake of my example it is important to recognize that the "flock" of parrots may be kept isolated from each other; the Baldwin effect depends on individual trainability and discernible differences therein, not on imitation or social learning, which are further effects of considerable power.

3. On the relation of unconscious and artificial selection (Darwin's terms) to natural selection, see Dennett (2001).

References

Avital, E. and E. Jablonka (2000). *Animal Traditions: Behavioural Inheritance in Evolution.* Cambridge: Cambridge University Press.

Baldwin, J. M. (1896). A new factor in evolution. *American Naturalist* 30: 441–451, 536–553. Reprinted in Belew and Mitchell (1996).

Belew, R. K. and M. Mitchell (eds.) (1996). *Adaptive Individuals in Evolving Populations: Models and Algorithms.* Reading, Mass.: Addison-Wesley.

Dawkins, R. (1982). *The Extended Phenotype.* Oxford: Oxford University Press.

Dennett, D. (1991). *Consciousness Explained.* Boston: Little, Brown.

Dennett, D. (1995). *Darwin's Dangerous Idea.* New York: Simon and Schuster.

Dennett, D. (2001). The evolution of culture. *Monist* 84 (3): 305–324.

Hinton, G. E. and S. J. Nowlan (1987). How learning can guide evolution. In Belew and Mitchell (1996), pp. 447–454.

Maynard Smith, J. (1987). Natural selection: When learning guides evolution. *Nature* 329: 761–762. Reprinted in Belew and Mitchell (1996).

Pepperberg, I. M. (2000). *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots.* Cambridge, Mass.: Harvard University Press.

Simpson, G. G. (1953). The Baldwin effect. In Belew and Mitchell (1996), pp. 99–110.

Spalding, D. (1873). Instinct with original observations on young animals. *MacMillan's Magazine* 27: 282–293. (Reprinted with an introduction by J. B. S. Haldane in 1954 in the *British Journal of Animal Behaviour* 2: 1–11.)