Dennett's Overlooked Originality

(Accepted by *Minds and Machines*: special issue on "Daniel Dennett and the Computational Turn")

David Beisecker
Department of Philosophy
University of Nevada, Las Vegas
4505 Maryland Pkwy, Box 455028
Las Vegas, NV 89154
phone: (702) 895-4038
fax: (702) 895-1279
email: beiseckd@unlv.nevada.edu

Abstract: No philosopher has worked harder than Dan Dennett to set the possibility of machine mentality on firm philosophical footing. Dennett's defense of this possibility has both a positive and a negative thrust. On the positive side, he has developed an account of mental activity that is tailor-made for the attribution of intentional states to purely mechanical contrivances, while on the negative side, he pillories as mystery mongering and skyhook grasping any attempts to erect barriers to the conception of machine mentality by excavating gulfs to keep us "bona fide" thinkers apart from the rest of creation. While I think he's "won" the rhetorical tilts with his philosophical adversaries, I worry that Dennett's negative side sometimes gets the better of him, and that this obscures advances that can be made on the positive side of his program. In this paper, I show that Dennett is much too dismissive of original intentionality in particular, and that this notion can be put to good theoretical use after all. Though deployed to distinguish different grades of mentality, it can (and should) be incorporated into a philosophical account of the mind that is recognizably Dennettian in spirit.

Dennett's Overlooked Originality

(Forthcoming in *Minds and Machines*)

1. Dennett's Philosophy: Positive and Negative Thrusts

There's no mystery why the *APA Committee on Computing and Philosophy* tabbed Daniel Dennett for the 2003 Barwise Prize. Few persons have done more over the past three and a half decades to extol the value of AI research in providing concrete thought experiments for epistemologists and philosophers of mind seeking to understand how our minds engage the world. And no philosopher has worked harder to set the possibility of machine mentality on firm philosophical footing. Dennett's defense of this possibility has both a positive and a negative thrust. On the positive side, he has developed an account of mental activity that is tailor-made for the attribution of intentional states to purely mechanical contrivances. According to this account, mental activity is best understood from the third person, quasi-engineering perspective of those who would *attribute* such activity to others. Individual mental states are understood as particular *phases* of abstract rational *patterns* of behavior, the discernment of which we find practically indispensable and which are more or less approximated by the various subjects we interpreting beings encounter around us.

Detractors have long argued that this third-person focus upon the *attribution* of intentionality is bound to miss something significant about our own inner mental lives. They complain that the relevant rational patterns of behavior *could* be observed in the absence of something vital – such as Searle's "intrinsic" intentionality or Block's "qualitative" or "phenomenal" dimension of consciousness. Dennett's staunch opposition to these critics is legendary. In his eyes, what allegedly goes missing – be it qualia, original intentionality, or even free will – is just so much philosophical mumbo-jumbo. Such notions are of little or no theoretical value, and should be dismissed as products of misguided and scientifically stultifying attempts to secure a special place for us with respect to the rest of nature. Thus he pillories any who would argue, typically from a first-person, phenomenological perspective, that there must be something special about us "genuine" intentional systems that elevates our mental capacities above the second-rate or "derived" intentional capacities of mere mechanisms and lower organisms. And in this connection, he is quick to remind us that we too could be regarded as sophisticated artifacts, designed as it were, or at least selected over eons, by evolution to exhibit behavior that for perfectly understandable reasons, turns out to be interpretable from an intentional stance. Given the gradual, incremental pace of these selective forces, whatever abilities we have, then, for intentionality, consciousness, and free will should be no different *in kind* from that already possessed in at least a rudimentary form by simpler systems: thermostats, chess-playing computers, the great host of other creatures – even zombies. And so we have the negative thrust of Dennett's defense of the possibility of machine mentality, wherein he ridicules as mystery mongering and skyhook grasping any

attempts to erect barriers to the conception of machine mentality by excavating gulfs to keep us "bona fide" thinkers apart from the rest of creation.

Dennett has largely gotten the best of these exchanges; for the most part, I think he's "won" the rhetorical tilts with his philosophical adversaries. Still, I worry that Dennett's "darker," negative side sometimes gets the better of him, and that this obscures advances that can be made on the positive side of his program. In this paper, I plan to show that Dennett is much too dismissive of original intentionality in particular, and that this notion can be put to good theoretical use after all.[1] Though deployed to distinguish different grades of mentality, it can (and should) be incorporated into a philosophical account of the mind that is recognizably Dennettian in spirit.

2. Original Intentionality: Dennett's Case Against

Dennett's published attempts to dispel the "myth" of original intentionality are not terrifically compelling. Rather than providing a blanket objection to the notion, he targets specific conceptions of original intentionality, which turn out to be saddled with unfortunate and unnecessary assumptions.[2] For instance, Dennett draws upon familiar *externalist* considerations in the philosophy of mind and language to show that the contents of subjects' mental states typically are not *intrinsic* to their internal physical or physiological constitution. From there, he argues that since meaning isn't wholly contained in the head, but instead may be determined by external features outside a subject's understanding, the contents of their intentional states are not *entirely up to those subjects themselves*, and so could not be appropriately *original*. Dennett's target here is evidently Searle's conception of "intrinsic" intentionality (or perhaps his unfortunate choice of that label). But even if we accept content externalism, this objection hits home only if the defender of original intentionality accepts the assumption that the contents of *originally* intentional states must be wholly determined inside the head. And it is simply unclear why one would have to make such a concession. Similarly, Dennett sometimes connects the issue of original intentionality with that of the alleged indeterminacy of translation.[3] If we cannot specify what even human subjects *really mean*, then how can we claim them to be *real believers*, possessing a kind of intentionality apart from the *merely attributed* or *derived* intentionality of artifacts? Once again, his thinking appears to be that if belief ascription is not wholly determinate, but rather depends partly upon hermeneutic choices made on the part of interpreters, then the intentional contents of a subject's beliefs could not be *wholly up to them* (or original). But just as one need not accept internalism about originally intentional states, it is also hard to see why a believer in original intentionality must be committed to the complete determinacy of originally intentional states. That intentional ascription generally admits of indeterminacies clearly does not show that our intentionality would have to be the same "derived" sort typically attributed to vending machines, chess-playing computers, and frogs.

Finally, Dennett sometimes suggests that his case against original intentionality turns on our evolutionary heritage. At this point, Dennett engages us in one of his trademark thought experiments. Suppose that we have designed robots that navigate their

surroundings with all the facility and mutual coordination as genuine human beings. Dennett urges us to accept that there is no scientifically well-grounded reason to believe that their intentional capacities would be any different from ours. On the further assumption that any designed artifact could only exhibit derived intentionality, it follows that our intentionality, too, could only be of the derived sort. But here it's unclear why a believer in original intentionality (at least one outside Searle's clutches) would have to embrace that further assumption. Granted, the intentional capacities of such envisioned robots might be equivalent to our own, but contrary to the second assumption, that only shows that original intentionality *could* be possessed by sufficiently fancy products of intelligent design. In any event, the thought experiment certainly doesn't show that there cannot be theoretically well-motivated distinctions in the kinds of intentionality exhibited by us and *simpler* intentional systems. Here, as elsewhere, we must be careful not to run different senses of "derived" together. To be sure, our intentionality presumably "derives" from Mother Nature in the sense that it is a *product* of natural selection, just as the intentionality of the envisioned robots would derive from us. But that doesn't mean that our intentionality must be derived in quite the same sense as that of simpler artifacts and organisms (we'll return to this sense later).

Still, it's clear that a large part of Dennett's refusal to embrace any notion of original intentionality stems from an overarching suspicion of cognitive saltations, fueled in turn by his conviction that the selective forces that shaped us brook few sharp distinctions. His assumption appears to be that original intentionality would have to be something that emerged suddenly, and that admits of no fuzzy boundaries. However, Dennett's reluctance to draw cognitive boundaries is itself remarkably selective. Despite his insistence on the continuity between lower and higher organisms, Dennett is not loathe to draw distinctions between their relative mental capacities. For instance, he sees significant differences among creatures that are and are not themselves capable of adopting the intentional stance, as well as between language users and the non-linguistic (reserving the capacity to form "opinions" to the former). And he sees fit to distinguish between those that are and are not able to reflect upon their own intentional states. Indeed, he suspects that qualiaphiles often misconstrue the latter distinction as that between us and zombies.[4] More significantly, Dennett is quite willing to classify creatures' mental capacities based upon their relative abilities to evaluate possible responses to situations, allowing them to kill off bad plans before those unwise courses of action wind up killing them.[5] Beginning with abjectly tropistic, or *Darwinian*, creatures (whose responsive dispositions evolve only through the operations of natural selection) his so-called "Tower of Generate and Test" ascends through *Skinnerian* creatures (whose responsive dispositions are modifiable through operant conditioning) and *Popperian* creatures (which are further capable of basing their behavior upon simulated outcomes of the possible responses that they might make). Finally, *Gregorian* creatures occupy the top of Dennett's hierarchy. These creatures are capable of designing tools of their own, including linguistic tools to discover ever better means of navigating their surroundings. As such, they are capable of jumpstarting a new, cultural form of evolution by serving as vehicles for "memes."[6] This final step, according to Dennett, "puts our minds on a different plane from the minds of our nearest relatives among the animals."[7]

As Dennett shows, creatures higher up the Tower of Generate and Test are better able to modify their behavior in order to adapt to their surroundings. Simply put, they exhibit greater behavioral flexibility and educability; the plans (or "future") that their minds produce die off in their stead. Nevertheless, Dennett resists any attempts to link these distinctions in educability to any alleged distinction in *intentional capacity*.

> My view is that belief and desire are like froggy belief and desire all the way up. We human beings are only the most prodigious intentional systems on the planet, and the huge psychological differences between us and the frogs are ill described by the proposed contrast between literal and metaphorical belief ascription. (*The Intentional Stance*, p. 112) [8]

More advanced forms of learning and adaptability are simply more sophisticated ways to support the same underlying kind of intentionality. Nowhere is this idea more evident than in his response to Dretske, where he argues that educability should not be regarded as any special mark of the intentional.[9] The capacity to learn from experience, whether it's individually or collectively instituted (that is, Popperian or Gregorian), is just one of several possible ways to get a subject to exhibit an overall rational pattern of behavior that is profitably viewed through the intentional stance. In principle, though it's much more cumbersome and demands more design foresight, one could also deploy natural selection or brute design to get a creature to exhibit a pattern of behavior that is similarly rational. So though there are many theoretically interesting differences in cognitive capacities, these differences are ill-suited to mark the alleged distinction between the relatively mindless and the genuinely rational.

Dennett's reasoning here is convincing only if one doesn't suppose that being educable (or capable of harboring memes) is an essential part of being appropriately rational. But that is precisely the idea that defenders of the educability standard for genuine mentality – folk like Dretske, Bennett, and even Davidson – are trying to articulate. Progress on this front is won only through a better understanding of what exactly constitutes an overall rational pattern of activity. However, Dennett's discussions of rationality are typically vague on this matter. I suspect that his attempts to debunk the notion of original intentionality are of a piece with a failure to consider the possibility that there might be *several varieties* of such patterns corresponding to different *types* of intentionality and different ways in which one can adopt an intentional stance. That is, I worry that Dennett treats his assumption of rationality as much too monolithic. And it is this monolithic view of rationality that causes Dennett to overlook what Dretske and Bennett are driving at: that educability – the ability to modify one's dispositions in the face of adverse experience – can be an essential facet of a distinctive kind of rational pattern.

So what are the essential elements that would render a pattern of activity rational? One thing Dennett does tell us is that in making the assumption of rationality required for adopting "the" intentional stance, interpreters expect subjects generally to act in their interests. But how do we interpreters determine what these interests are, or what it is subjects *should* do? Where does this normative element come from? In the case of the abjectly tropistic (e.g., the *sphex* wasp or the frog of philosophical legend, who famously stuffs itself silly with lead pellets), we identify what a subject should do in terms of what it has been selected or designed to do – that is, its natural (biological) or selective purposes.[10] Failures on a subject's part to carry out those purposes are generally explained in terms of *kludges*, design shortcuts, or failures on the part of a designer to

anticipate the kinds of situations its creation will face. It's this very identification of a creature's interests with those of its designers (or it genes), along with the concomitant inclination to blame lapses of rationality on those designers (or Mother Nature) that accounts for the idea that the intentionality of relatively tropistic creatures is somehow derived or second-rate. We are apt to discount froggy behavior as manifesting true belief because it seemingly fails to respond rationally to its mistakes, or to respond to them *as mistakes*. One is tempted to say that it isn't capable of getting things right or wrong "by its own lights." This suggests that an *original* sort of intentionality would have to include some measure of *self-correction* in order to capture behavior appropriately governed by a subject's own *acknowledgement* of the norm of correctness. But in order to sustain the claim that such activity is appropriately regarded as involving the *correction* of *errors*, it would seem that we must have some account of a subject's aims as well, for how could we recognize mistakes *as mistakes* unless they are somehow liable to prevent a subject from attaining its desired ends? That is, it would be difficult to tell a story with the requisite normative punch without including some account of goals. It is thus reasonable to suppose that discernibly rational activity requires elements of both critical (self-corrective) and practical (means-end) reasoning. So to a first approximation, I propose we regard a pattern of activity as discernibly rational if it exhibits self-corrective behavior that is directed towards some goal. With that in mind, in the next section I will show that the activity of certain educable creatures is discernibly rational in this more precise sense.[11] They can exhibit a distinctive type of goal-directedness, where the goals in question aren't so tightly to a subject's selective purposes, and indeed, can be discerned independently of them.

3. Expectation-Mongering: Rationality in Education

As mentioned above, Dennett's objections to original intentionality are piecemeal, and leave open the possibility for one to defend a conception of the notion that is free of the problematic assumptions Dennett associates with appeals to original intentionality. In the remainder of this paper I propose to construct an account of a relatively simple *kind* of original intentionality, which does just that. Let's begin with ethology. To account for certain *blocking effects*, several learning theorists have argued that the observed educability of some animals is best explained in terms of the adjustment of *expectation-like* structures mediating between sensory input and behavioral output.[12] Accounts of expectation-based educability aim to capture (that is, to describe in suitably informative terms) patterns of activity, whereby creatures exhibit an apparent sensitivity to the *consequences* of their own responses. The basic idea is that the actual responses such creatures make in situations is a function of various outcomes that they currently associate with the particular responses available in their behavioral repertoires, outcomes whose associations may change over the course of a creature's experience. Since different responses in the same situation can bring about different outcomes, and since the same type of response can, depending upon the circumstances, yield different outcomes, the abstract "expectation" structures posited to mediate between sensory input and behavioral output need to include (at least) three separate components: 1) conditions of activation (and deactivation), 2) a response type, and 3) a consequence condition. The

first component specifies, as it were, when an individual expectation is turned "on" and "off". When an expectation is activated (or "on"), the creature associates the outcome specified by that expectation's third component with the response specified by its second.[13] Should the creature engage in that response and the consequence condition *not* be satisfied, then the creature would be disposed to adjust the components making up that expectation. Through the revision of expectations when they are so "violated" these creatures distinguish themselves from the brutely tropistic, and so display the sensitivity to the consequences of their own responses that learning theorists have sought to describe.[14] So the story is basically this: under certain circumstances, an expectation will be activated and the creature will then anticipate that a certain response will yield a particular outcome. Should that turn out not to be the case, its dispositions to form such anticipations will change. In true Dennettian fashion, then, we can identify *expectation-mongering creatures* as those whose overall pattern of behavior is most systematically and fruitfully described as governed in part by the adjustable anticipations it has of its various responses, which are generated by an evolving stock of expectations.[15]

So construed, expectations are abstracta; they help to characterize a creature's overall pattern of behavior while remaining silent about the specific physiological structures inside their brains. Note also that the description I gave of expectation-mongering behavior doesn't presume any antecedent understanding of a creature's goals or purposes. In particular, it doesn't make any obvious appeal to the purposes for which a subject has been designed or selected. One can identify expectation-mongering creatures as such without having to recognize them as products of design or subject to selective pressures. Nor have I construed expectation-based educability as the selection of responsive dispositions that have positive survival value, although that is presumably something such behavioral plasticity can bring about. So given the intuition that we ought to be able to evaluate expectations as correct or mistaken (and so contentful), this would seem to be a promising beginning of a story about a type of intentionality that doesn't depend upon (or derive from) the identification of a creature's selected purposes. However, to sustain this claim, we need to show in detail how the behavior of expectation-mongering creatures fits an overall *rational pattern*. That is, we need to show how expectation-mongering can be viewed as goal-directed, self-corrective activity. While it's fairly intuitive how this story should go, the details can be a bit tricky. So bear with me; my strategy will be to begin with the practical side, and construct an account of *goals* from this account of expectation-mongering, and then turn around and use this account of goals to ground the notion of *expectation error*.

One especially nice thing about starting with expectations is that it affords a straightforward and satisfying specification of goals.[16] A certain outcome is to be regarded as one of a creature's current goals, to the extent that the creature is disposed to engage in responses *expected* to bring about that outcome.[17] A creature that is systematically disposed to engage in responses associated with the outcome of, say, acquiring cookies can be understood as having the acquisition of cookies as a goal. Another creature, disposed *not* to engage in responses associated with an electric shock, can be understood as having an *aversion* to shocks. Like expectations, goals are abstract posits, which work work in conjunction with a creature's expectations to make sense of

its particular responses to situations. By characterizing responses in terms of the outcomes they are expected (by the subject) to bring about, these explanations show how a particular response fits a creature's overall pattern of responsive dispositions. And we need not regard such explanations as empty, because they point out that a subject might have done otherwise, had that response not been expected to bring about a certain outcome, or had some other response been expected to bring about that outcome instead. Notice in particular that an expectation must have an appropriate consequence condition before it can be paired with a goal in order to explain a creature's behavior. The expectation's *content* – as given by its consequence condition - must itself satisfy the goal's *condition of satisfaction*.[18] Since goals and expectations must have the right sort of "fit" with one another before they can successfully explain a creature's behavior, these explanations face what could be thought of as a *rational constraint*. Thus it makes some sense to claim that attempts to explain a creature's behavior with respect to its goals and expectations to be attempts to *rationalize* its behavior.

While this is obviously a broadly dispositional account of goals, it's worth noting that it does not crudely identify a creature's goals with the outcomes that the creature is actually likely to bring about.[19] On this proposal, creatures do not have to be disposed to bring about the eventual attainment of their goals. For one thing, just as we can pick out fragile objects without requiring that they manifest their fragility by shattering, we can identify a creature's goals, even though it might not ever find itself in circumstances where their attainment is possible. For our purposes, however, the respects in which the activation of *expectations* can block the attainment of goals are particularly significant. Here we can say that an expectation-mongering creature will be disposed to attain its goals (whenever such attainment is possible) to the extent that its expectations are configured *correctly*.[20] This, of course, is where the normative rabbit gets pulled out of the naturalistic hat. The nice thing is that we can pick out unfavorable expectation configurations likely to hinder a creature's attainment of its goals, and so have reason to regard these configurations as expectation *errors*. Naturally enough, a creature is liable not to fulfill a goal if one of its expectations is activated in a situation in which the expectation's response would fail to bring about the satisfaction of its consequence condition. A creature is likely not to fulfill its goal of acquiring cookies if a response it associates with the outcome of acquiring cookies will actually bring about a different outcome instead. We can thus think of such an occurrence as an *error of commission*. Similarly, *errors of omission* arise whenever the response of an expectation that is *not* activated would bring about the satisfaction of its consequence condition (that is, were its activation not to be an error of commission). Here our creature is liable not to engage in a response that would procure cookies, since it fails to associate that response with that desired outcome. Since these two expectation configurations are liable to prevent a creature from attaining its goals, expectation-mongering creatures are susceptible of two distinct sorts of mistakes about the way things are in their environments.[21] They can be evaluated as having gotten things right or wrong, and so can be understood to exhibit a type of intentionality above and beyond that typically attributed to artifacts and simple organisms. Observe once more that while expectations are, as it were, *ontologically* or conceptually prior to goals in the sense that the latter can only be defined in terms of an antecedently intelligible account of the former, goals nevertheless enjoy a *normative* priority over expectations in the sense that

the notion of expectation *error* depends upon (or is intelligible as such only with respect to) this account of goal-directedness. While goals owe their *existence* to expectations, expectations owe their *normativity* to goals.

So we now have found reason why, from a creature's own perspective, its expectations *ought* to be activated just in case their consequence conditions would be satisfied, were the creature to engage in the response picked out by that expectation's response component. As an account of error, this story has several appealing features. Heading that list is the fact that, unlike "teleobiological" accounts of intentionality that appeal to design or proper functioning, the commission of these errors doesn't depend upon any antecedent determinations of when given responses tend to have survival value or to be reproductively advantageous for a creature. In addition, these standards for expectation correctness are *categorical* in the sense that they apply as they do, irrespective of the particular goals a creature might possess. The activation (or inactivation) of an expectation can be identified as correct or mistaken, regardless of what a creature's goals happen to be. Moreover, the conditions for the appropriate activation of one expectation can be quite different from the conditions of appropriate activation for another. That is, the activation of separate expectations can be beholden to distinct features of a creature's environment. As a result of this *feature selectivity*, expectation-mongering creatures can be correct with respect to some features of their environment, yet mistaken with respect to others. They can get things right or wrong *in a variety of respects* due to the simultaneous activation of several expectations. In fact, an expectation-mongering creature could even be *massively* mistaken about the way things are.[22] Furthermore, the situations in which one expectation would be appropriately activated might just happen to line-up or co-vary with those in which another would be activated. For instance, the circumstances in which one response would procure cookies might be precisely those in which another response would bring on an electric shock. "Extensionally speaking," distinct expectations can thus share the same circumstances of appropriate activation. However, the *particular means* by which these circumstances are picked out would differ for each such expectation on account of their differing expectation components. So even though their circumstances of appropriate activation can be the same, their content ("intensionally speaking") can remain quite distinct. Had the subject's environment been otherwise, these expectations might not have shared circumstances of appropriate activation. It would thus appear that attributions of expectation states exhibit something like the ballyhooed semantic opacity or sensitivity to intensional contexts so often associated with the attribution of genuine intentional states. To attribute an expectation to a creature is not tantamount to attributing to it other expectations sharing the same circumstances of appropriate activation.

It should also be evident how expectation-mongering creatures can be understood as exhibiting a certain measure self-corrective, *critical* rationality as well as the practical rationality I've just described. Insofar as they are disposed to revise their expectations in the wake of the errors described above, such educable creatures take discernibly rational steps to minimize future mistakes. Of course there's no guarantee that these revisions will yield future success.[23] The point is just that creatures displaying this sort of educable capacity would take expectation correctness or aptness to be a *regulative ideal*, at least in

the sense that they are disposed to revise error-prone expectations while leaving correct expectations as they are. And so it seems that they display something akin to rational responsiveness to error that Davidson argues must be possessed by any rational animal.[24] By responding in a more or less reasonable fashion when the outcomes of their responses aren't as they were expected to be, such creatures manifest an apparent capacity to be "surprised." In sum, then, it should be clear how expectation-mongering creatures can exhibit a rational pattern of self-corrective goal-directed behavior that is wholly indiscernible in the behavior of the abjectly tropistic.

4. Conclusion: Original Intentionality, Dennett Style

In the previous section, I took pains to show how the behavior of a certain sort of educable creature can exhibit a distinct type of rational structure. As I've shown, this type of behavior supports the attribution of primitive doxastic and conative states (expectations and goals), which can be identified without any obvious appeals to natural purposes or proper functioning. So I conclude that I've succeeded in profiling a *kind* of intentionality that is not derived from designed purposes in the same way that the intentionality of simpler beings is, and which could have emerged gradually through evolution. It would seem reasonable to call such intentionality "original;" or perhaps it would be better to call it *sui generis* – that is, of its own kind. And there is no reason to suppose that some of our more sophisticated artifacts couldn't exhibit this type of intentionality as well (thus defusing one of Dennett's objections to original intentionality). Indeed, since expectation-mongering creatures wouldn't have to be products of *any* sort of selection, natural or otherwise, and their goals and expectations are intelligible as such without our having to consider the purposes for which they have, as it were, been designed, this account shows how non-biological "creatures"- for instance, those philosophical fantasies spontaneously generated out of swamp muck - could nevertheless possess this kind of intentionality. Moreover, not only is this account of goals intelligible apart from considerations of the purposes for which a creature has been designed or selected, these goals might even collide with those purposes. For instance, there is no reason why a creature couldn't be disposed to respond in ways expected to bring about reproductively disadvantageous outcomes. Such a creature would have a goal that, from a biological point of view, is remarkably maladaptive.

Now it is worth pausing to review how thoroughly Dennettian in sprit this account of original intentionality is. To begin with, it focuses upon *patterns* of behavior discernible by adopting a certain kind of interpretive stance. The contentful, intentional states (the goals and expectations of expectation-mongering creatures) that make rational sense of a subject's behavior are abstracta, which are unintelligible apart from these overall patterns. As such, they are not the discrete inner physical states of the kind preferred by "industrial strength" realists like Fodor, but rather are like the abstract posits favored by Dennett's own brand of mild realism. Moreover, since these rational patterns might only be approximated in actual subjects, there is plenty of room for creatures to "more or less" exhibit original intentionality. As a result, this account readily allows for fuzzy boundaries between those with and without original intentionality.

Perhaps more interesting, however, is that there is no claim that the story would have to end right here – that the kind of intentionality discernible in expectation-mongering must be that to which our own high-grade intentionality can be assimilated. We might be able to profile more sophisticated kinds of rational patterns from which new types of intentionalities emerge. To be sure, we linguistic creatures evidently engage in performances that can be evaluated as true or false (and otherwise appropriate or inappropriate) according to norms that are instituted across our linguistic communities. As such, we can be held to rationality standards that are not binding on the non-linguistic. The above story about expectation-mongering would seem far too atomistic and individualistic to account for these peculiarly discursive and social forms of intentionality. Moreover, the kinds of goals I've identified in expectation-mongering are not "original" or "wholly up to a creature," in the sense that they are subject to its own choices. However, the fact that we can describe in broadly naturalistic terms a pattern of activity in which an (albeit primitive) *original* type of intentionality can be discerned should give us hope that we can describe other patterns in which more sophisticated kinds of intentionality can be descried, including the irreducibly social stripe that we enjoy. Indeed, the story just told could prove to be a good *platform* upon which to erect *further* accounts that set more involved intentional capacities as *targets*.[25]

This story, then, has a liberating moral. Many researchers tacitly buy into a monolithic view of intentionality, according to which ultimately there is but one correct way to draw "the" boundary between *true believers* and systems with *merely derived intentionality*. Such thinking only encourages the idea that by telling *one* story about how to discern intentional states in subjects that one has thereby told them *all*, and that rival ways of drawing the distinction between those with and those without minds must be mistaken. That attitude only fosters interminable debates among philosophers, psychologists, ethologists, and AI researchers over competing standards of mentality: language use, second-order intentionality, or the expectation-based educability profiled above. Dennett's himself has tried to rise above this fray by denying that there is any such boundary to be drawn; there's only a smooth continuum stretching from simple intentional systems, like frogs and thermostats, to us, the most prodigious intentional systems on the planet. In so doing, Dennett himself subscribes to a wildly permissive (hence oft-criticized), yet equally monolithic view of intentionality. But all of that can (and should) be avoided. The preferable, and ultimately more perspicuous, way to defuse the debate is to admit of *several* intelligible distinctions in kinds of intentionality, corresponding to different ways in which interpreters can discern rationality in the behavior of subjects. From the fact that there is no single, privileged mark of the mental it certainly doesn't follow that there are no well-motivated distinctions in intentional capacities altogether. One shouldn't deny original intentionality, but rather embrace it in its myriad forms.

References

Bennett, J. (1990), *Linguistic Behavior*, Indianapolis, IN: Hackett.

Bennett, J. (1991), 'Folk-Psychological Explanations', in J. Greenwood, ed., *The Future of Folk Psychology: Intentionality and Cognitive Science*, Cambridge: Cambridge, pp. 176-195.

Bermudez, J. (2003), *Thinking Without Words*, Oxford: Oxford.

Dennett, D. (1987), *The Intentional Stance*, Cambridge, MA: Bradford/MIT.

Dennett, D. (1988), 'Quining Qualia', in Marcel and Bisiach, eds., *Consciousness in Contemporary Science*, Oxford: Oxford.

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

Dennett, D. (1996), *Kinds of Minds*, New York: Basic Books.

Dennett, D. (1998), *Brainchildren: Essays on designing minds*, Cambridge, MA: MIT.

Dickinson, A. (1989), 'Expectancy Theory in Animal Conditioning', in Klein and Mowrer, eds., *Contemporary Learning Theories*, Hillsdale, NJ: Lawrence Erlbaum.

Dretske, F. (1988), *Explaining Behavior*, Cambridge, MA: MIT.

Dretske, F. (1999), 'Machines, Plants, and Animals', *Erkenntnis* 51, pp. 19-31.

Heyes, C, and A. Dickinson. (1990), 'The Intentionality of Animal Action', *Mind & Language* 5, pp. 87-104.

Levy, D. (2003), 'How to Psychoanalyze a Robot: Unconscious Cognition and the Evolution of Intentionality', *Minds and Machines* 13, pp. 203-212.

Russell, B. (1927), *The Analysis of Mind*, Allen and Unwin.

Staddon, J.E.R. (1983), *Adaptive Behavior and Learning*, Cambridge: Cambridge.

And 3 articles by the author

[1] Though this paper focuses on original intentionality, I would argue that Dennett is also too dismissive of the idea of *qualia*. Elsewhere [2004], I've shown how to develop an account of the qualitative dimension of conscious experience by making reasonable and intuitive sense of both philosophical and non-philosophical *talk about* "what it's like" to have certain conscious experiences, including most importantly, attributions of *knowledge* of what it's like to have such experiences. On this account, the explanatory gap appears because such talk plays a perfectly unmysterious *epistemic* function that cannot be played by physical or scientific vocabulary. Even though it flies in the face of claims Dennett makes about the notion

of qualia (1988, as well as chapters 7-11 of *Brainchildren*), this account can be regarded as defending a "mild realism" about qualia on the model of Dennett's own "mild realism" about belief and intentional notions more generally. Since the primary data to be accounted for are the things we are inclined to *say* about conscious experience (both our own and that of others), such an approach accords well with the "heterophenomenological method" Dennett himself recommends. In any event, there is nothing in it to contravene Dennett's rejection of the Cartesian Theatre in favor of the multiple drafts model of consciousness, which has more to do with the putative distinction between conscious and unconscious mental states than with the existence of qualia.

[2] Among other places, Dennett discusses these matters at length in chapter 8 of *The Intentional Stance*, chapter 14 of *Darwin's Dangerous Idea*, and chapter 2 of *Kinds of Minds*. For a more complete rejoinder, see the author's [2001]. The issue has also recently graced the pages of *Minds and Machines* (see Levy, 2003)

[3] See, for instance, "Real Patterns" (reprinted as chapter 5 of *Brainchildren*, pp. 114ff.). Dennett typically uses indeterminacy considerations to argue against "industrial strength realism" in favor of his "mild realism."

[4] See *Brainchildren*, Chapter 10 ("The Unimaginable Preposterousness of Zombies").

[5] See, for instance, *Darwin's Dangerous Idea*, Chapter 13, and *Kinds of Minds*, Chapter 4.

[6] With this new cultural selection in place, Dennett tells us that interpreters can begin to see new, "memetic" points of view arising, which might come into competition with those of our genes. I find it striking that with all this proliferation of points of view, Dennett seems to neglect *our very own individual* points of view, as distinct from those of our genes or memes, and possible coming into conflict with them. Doing so, however, would seem to acknowledge an original form of intentionality, since of course genes and memes, unlike us, aren't *really* the kinds of things to have points of view.

[7] *Darwin's Dangerous Idea*, p. 381

[8] Consider as well the last paragraph of "Do Animals Have Beliefs?" (*Brainchildren*, Chapter 22, p. 331): "We find that there are many, many differences [in cognitive abilities], almost all of them theoretically interesting, but none of them, in my opinion, marking a well-motivated chasm between the mere mindless behavers and the genuine rational agents." These claims are all the more striking when you consider Dennett's own illustrations of his various creature-types (*Darwin's Dangerous Idea*, 374-5), which depict his higher level creatures as themselves forming *pictures* of the world in some inner environment. Now what could that possibly represent, if not some special intentional capacity?

[9] See *Brainchildren*, Chapter 3 ("Do-It-Yourself Understanding").

[10] Dennett seems to think that considerations of design or selected purposes are endemic to *all* adoptions of the intentional stance, an attitude which is only fostered by his understanding of *homuncular functionalism*: the idea that complex intentional systems are to be regarded as composed of simpler systems, which themselves turn out to be interpretable by "the" intentional stance. Hence he occasionally claims that the intentional stance is a limiting form of the design stance. See, for instance, "Cognitive Ethology: Hunting for Bargains or a Wild Goose Chase?" (*Brainchildren*, Chapter 21, p. 312).

[11] To be sure, several philosophers have argued that educability marks an important distinction between the intentional capacities of creatures. See in particular Dretske (1988, 1999) and Bennett (1990, 1991). Unfortunately, these discussions typically dwell upon how educable capacities render organisms better able to fulfill their natural purposes in the face of environmental contingency. Dretske, for instance, focuses on how providing creatures with the ability to conduct their own selection of appropriate internal indicators might be the best way for a designer (including Mother Nature) to solve the problem of constructing creatures that are likely to fulfill their intended purposes. I would argue (and Dennett likely agree) that this focus upon solving the "design problem" renders the account ill-suited to describe the distinction between derived and original intentionality. If, however, we had an independent story about how the flexibility of educable creatures gives rise to a *sui generis* sort of intentionality, then we might begin to see how an *original* intentionality could emerge gradually as a product of natural selection.

[12] See, for instance, Staddon (1983, pp. 414ff.) and Dickinson (1989). For example, some animals that have been trained to associate a conditioned stimulus with an unconditioned stimulus will subsequently fail to associate other stimuli with the unconditioned stimulus, when the latter are presented along with the original conditioned stimulus. Rats that have been trained, for instance, to associate a bell tone with an electric shock will not come to associate a red light with a shock, as long as the red light is consistently paired with the bell tone. The prior conditioning prevents (or "blocks") subsequent conditioning to other,

co-varying stimuli. If this learning were merely a matter of the frequency of stimulus-pairing, then one would expect the animal to become conditioned to the new stimulus as well. One would expect the rats eventually to associate the red light with a shock, as indeed they do when they aren't subjected to the earlier training. Many learning theorists have argued that the failure of previously conditioned animals to become conditioned to the new stimulus arises because the animal already uses the original conditioned stimulus to *predict* the occurrence of the unconditioned stimulus, and with a reasonable degree of success. When a previously conditioned rat encounters the compound tone and light stimulus, it *expects* that the shock will occur (because it heard the bell tone), and so the subsequent shock isn't a *surprise*. Since events are as they were *expected* to be (they were not novel), there is no pressure to develop new associations, and there is no subsequent conditioning to the light. Thus these theorists conclude that the rats are responding to *surprise*, to things not being as they *expected* them to be.

[13] To be relentlessly naturalistic, the first and third components could be specified in terms of, say, activity along a creature's sensory manifold, while the second in terms of the activation of particular motor programs.

[14] Different accounts of expectation-based educability differ with respect to which expectation components are allowed to vary from expectation to expectation, which components are capable of being altered, and also the conditions in which they stand to be adjusted. Staddon (1983), for instance, takes learning to involve the adjustment of consequence conditions, while Bennett (1990) effectively restricts it to the revision of activation conditions. A fully general account of expectation would leave as much of this up for grabs as possible.

[15] More formally: their response to situations is a partial function of the consequence conditions of currently activated expectations. It bears mentioning that I'm not trying to show that any particular creatures are expectation-mongerers, which would be the work of ethologists, not philosophers. Notice also that I've defined expectation-mongering in terms of how a creature *would behave* in various possible situations. Since any pattern of actually observed behavior could be the product of tropisms, showing that a creature is an expectation-mongerer would have to involve establishing that certain counterfactuals hold. It turns out, then, that those who *design* devices would likely have an easier time justifying the attribution of expectations to their subjects than those who encounter them "out in the field," simply because they have a better sense of what goes on inside the "black boxes" they investigate, and so would have a better grasp of the relevant counterfactuals. For a discussion of the difficulties attributing to wild subjects states similar to the expectations described here, see Heyes and Dickinson (1990).

[16] Insofar as the goals so-construed rest upon an antecedently intelligible account of expectation, this account reverses the strategy historically advocated by Ramsey, and most recently pursued by Bermudez (2003).

[17] In a similar fashion, we can determine a preference *ordering* among outcomes. Curiously, such an ordering might even turn out to be pairwise intransitive. Such apparently irrational goal structures should serve to remind us that not all expectation-mongering creatures need to be understood as having intelligible goals.

[18] Note that such contents may be specified in *either* proximal or distal terms, depending upon the pragmatic interests of interpreters. Such latitude doesn't mean that "anything goes" in content specification, nor does it impugn its status as *originally* contentful.

[19] It's also worth remarking that this account of goals is not a "reinforcement" theory, such as that defended by Dretske (1988) and Bermudez (2003). Nor is it an "extinction" theory, such as that occasionally attributed to Russell (1927).

[20] As Bennett (1990, pp. 42ff) might claim, a creature will be disposed to attain its goals "all things being equal," where having correctly configured expectations is a crucial part of things being equal.

[21] Please observe that this account doesn't rule out *accidental* (or unexpected) success at attaining goals. Indeed, it makes the notion of accidental attainment intelligible.

[22] That is, the so-called principle of charity need not apply to the ascription of a creatures expectations (though it still might apply to more linguistically infected intentional states).

[23] Against the background of this account of expectation error, we can further understand expectation-mongering creatures to be making *errors of expectation revision* whenever they adjust an expectation in ways that would render it more susceptible to either errors of commission or errors of omission.

[24] To be sure, Davidson further tries to argue that the conceptual resources required to be surprised in turn require an animal to be capable of *interpreting the utterances* of others; thus thought requires talk.

However, we don't have to accept this additional claim to take Davidson's point that the capacity to be surprised, or to recognize when the way things are aren't as one took them to be, is an important part of being a rational animal. For more discussion and criticism of Davidson's position, see the Author's [2002].

[25] For instance, we can begin to capture primitive *inferential* capacities by expanding the activation (and deactivation) conditions of expectations to include not just immediate sensory conditions, but also the activation (or deactivation) of other specific expectations. Creatures would then be able to assemble their expectations into primitive networks governed by entailment and exclusion relations. I also have some ideas about how to capture primitive forms of *linguistic meaning* – thoughts which draw from the broadly pragmatist idea that the meaning of a sign is some function of its *expected* consequences (see [forthcoming]).