Darwin and Dennett: The Operationalist Debate and the Teleological Response

Jonathan Roorda
Massachusetts Institue of Techonology

"What, Klapaucius, would you equate our existence with that of an imitation kingdom locked up in some glass box?!" cried Trurl. "No, really, that's going too far! My purpose was simply to fashion a simulator of statehood, a model cybernetically perfect, nothing more!"

"Trurl! Our perfection is our curse, for it draws down upon our every endeavor no end of unforeseeable consequences!" Klapaucius said in a stentorian voice. ". . . Don't you see, when the imitator is perfect, so must be the imitation, and the semblance becomes the truth, the pretense a reality!. . ."

"Sheer sophistry!" shouted Trurl, all the louder because he felt the force of his friend's argument. ". . . The subjects of that monster Excelsius do in fact die when decapitated, sob, fight, and fall in love, since that is how I set the parameters, but it's impossible to say, Klapaucius, that they feel anything in the process — the electrons jumping around in their heads will tell you nothing of that!"

"And if I were to look inside your head, I would also see nothing but electrons," replied Klapaucius. ". . . You say there's no way of knowing whether Excelsius' subjects groan, when beaten, purely because of the electrons hopping around inside — like wheels grinding out the mimicry of a voice — or whether they really groan, that is, because they honestly experience the pain? A pretty distinction, this! No, Trurl, a sufferer is not one who hands you his suffering, that you may touch it, weigh it, bite it like a coin; a sufferer is one who behaves like a sufferer!"

This dialogue from Stanislaw Lem's charming collection entitled *The Cyberiad: Fables for the Cybernetic Age*, captures perfectly the nature of a debate which has raged for nearly four decades among philosophers, psychologists, and computer scientists. Like Klapaucius' and Trurl's argument,

this debate focuses on the precarious status of the inner mental life of human artifacts which exhibit certain aspects of convincingly human behavior. In the real world, however, the artifacts in question are digital simulations not of kingdoms but of individual minds, and they are instantiated not in glass boxes but in the computers which have become such a familiar presence in modern society. This debate over the possibility of expressing true intelligence in terms of a computer program has found its two most eloquent rivals in the legendary war-era British computer scientist Alan Turing and the tenacious Berkeley philosopher John Searle. Their respective papers on artificial intelligence form the antipodal landmarks around which the rest of the debate has been mapped. Fundamentally, however, the differences between Turing and Searle reflect not only upon the specific issues of machine intelligence but upon more basic philosophical and scientific questions which can be traced back to the eighteenth century and to the question of the existence of purpose in the world of natural creation. Like the artificial intelligence debate, this issue had two definitive antagonists, David Hume and William Paley. The intellectual conflict surrounding their works extended into the nineteenth century, when Charles Darwin published his seminal *Origin of Species.* Like many other debates, the issue of purpose was derailed by the upheaval which followed Darwin's work, as its fundamental assumptions where called into question and eventually fused into the Darwinian synthesis. Today, a new intellectual synthesis seems to be forming, and the antipodes of Turing and Searle are drawn closer together by en inchoate philosophical tradition inspired by Daniel Dennett. Borrowing a page from Darwin, Dennett simultaneously reconciles and dismantles the arguments of Turing and Searle, using precisely the same philosophical mechanism by which Darwin both vindicates and undermines Hume and Paley.

Although the idea of artificial intelligence as a serious conceptual possibility dates back to Charles Babbage, its first coherent philosophical expression is found in Alan Turing's "Computing Machinery and Intelligence," published in 1950. Turing, at the time a prominent although socially ostracized figure in the developing field of computer science, turns his attention to the question, "Can machines think?" He quickly rejects this formulation of the question as incoherent, pointing out that it contains terms whose extensions are too vaguely defined to be pressed into reputable philosophical service. In his own words, "The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the

words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?' is to be sought in a statistical survey such as a Gallup poll." Turing insists that the question must be replaced with a second formulation, one which avoids conceptual terms shaded with nuances of interpretation and which relies only on well-defined, observable phenomena. From this basic operationalist stance, Turing proceeds to define his famous Turing test, according to which a human observer interrogates two hidden conversationalists and attempts to ascertain which is a human and which is a computer. The question which Turing considers can now be formulated as "Can a computer be programmed to pass the Turing test?", and Turing devotes the rest of the article to defending his argument that the two questions can be substituted for one another against a wide range of objections.

An important point which must be recognized is that Turing does not offer his article as a defense of the ultimate possibility of artificial intelligence. At one point, he surmises that computers with a storage capacity of one thousand megabytes will be able to pass the Turing test by the end of the century; however, he offers no arguments to support this conviction, and he abandons it as merely a tangential point in his essay. As he admits, "The only really satisfactory support that can be given for the view expressed [in favor of artificial intelligence] will be that provided by waiting for the end of the century and then doing the experiment described." Instead, Turing seeks to formulate a criterion which can be used to arbitrate the emotionally heated arguments surrounding artificial intelligence in a systematic way. He is less interested in defending the pursuit of artificial intelligence than in devising a mechanism to judge the products of that pursuit. In taking up this challenge, Turing finds himself confronted with the same dilemma which haunted the behavioral psychologists of his day: the seeming necessity of defining mental phenomena in purely observational terms. Turing correctly realizes that an intellectual consensus on machine intelligence can never be reached by appealing to the wildly varying institution which exist on the nature of intelligence; agreement can only be reached by reducing the question to one which can be answered through appeal to accessible, reproducible data. The strength of the Turing Test is that it reformulates the questions of artificial intelligence in a way that simultaneously appeals to our intuitions of linguistic behavior as an exclusive product of human-like intellect, preserves the vagueness inherent in the original

question by relying on the judgment of an interrogator, and utilizes a controlled set of experiments with verifiable results. In introducing this mechanism, Turing violates an unspoken philosophical tradition by insisting that our intuitions be forced to conform to our rigid conceptual formulations, rather than the other way around. Since "at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted," it makes no sense to base philosophical arguments on such malleable intuitions. From Turing's standpoint, the resolution of the artificial intelligence question must derive not from idle speculation concerning the ultimate nature of mental phenomena but from the establishment of a rigid scientific yardstick which permits no ambiguities in verification.

For John Searle, on the other hand, such a yardstick can only be an obfuscating device used to promote a degenerate ideology which has captured the minds of computer scientists. In his 1980 paper "Minds, Brains, and Programs," he mounts a scathing attack against "the claims I have defined as those of strong AI, specifically the claim that the appropriately programmed computer literally has cognitive states and that the programs thereby explain human cognition." Although he cites the contemporary programming work of Schank, Winograd, and Weizenbaum as primary targets of his critique, the argument which he develops seems aimed directly at Turing, whom he views as the primary godhead of the artificial intelligence pantheon. Searle's infamous "Chinese room" thought experiment attacks the premises of the Turing test by constructing a hypothetical mechanism, analogous to a digital computer program, which is able to pass the test and yet which seems intuitively to fall short of any reasonable standard of human intelligence. Searle imagines himself confined to a room along with a huge body of uninterpreted Chinese characters and a comprehensive set of formal rules for their syntactic manipulation. Native Chinese speakers pass messages written in Chinese to him; he applies the algorithm to these messages and returns the resulting character sequences, which are actually appropriate responses in fluent Chinese. This system represents a finite program which could in theory be instantiated on a digital computer and which would presumably be able to pass any Turing test administered by a Chinese speaker. Yet, as Searle writes, "it seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker,

and I can have any formal program you like, but I still understand nothing." In other words, Searle argues that the Turing test must fail as a criterion of intelligence, because his hypothetical computer program simulates the external linguistic behavior of human intellect in every particular, yet lacks understanding, a crucial factor in any conception of inner mental life.

Now, the battle lines are drawn between Searle and Turing, and it is worthwhile to reflect on their points of similarity and difference. Both Turing and Searle agree that human beings engage in certain behavioral patterns as a direct consequence of their possession of a set of faculties and inclinations which are collectively referred to as "intelligence"; in addition, both agree (at least for the sake of argument) that it is in theory possible to program digital computers to produce behavior which is identical to intelligent human behavior in all relevant aspects. They differ in their beliefs on what these two facts imply. For Turing, the fact that machines can instantiate intelligent behavior proves that they are at least in principle capable of intelligence in the full sense defined above. Turing argues from a position which Searle dismisses as "residual behaviorism or operationalism": the position that concepts such as intelligence are coherent only when defined in terms of the observable phenomena by which they are characterized, and thus that whatever produces these observable phenomena falls completely within the scope of the concept. Stripped of the emotional baggage it has acquired in recent philosophical and psychological discourse, the term "operationalism" seems a good one to use to refer to Turing's essential stance. Searle, on the other hand, opposes operationalism in all its forms. From his perspective, even though intelligence is ultimately defined in terms of unobservable "causal powers" which cannot be instantiated through any level of syntactic manipulation and whose presence, although presumably impossible to verify experimentally, nevertheless serves as an absolute requirement for the existence of true intelligence. Searle illuminates this position when he considers the natural tendency to attribute intelligence to any source of intelligent behavior: "The reason we make these attributions is quite interesting, and it has to do with the fact that in artifacts we extend our own intentionality; our tools are extensions of our purposes, and so we find it natural to make metaphorical attributions of intentionality to them; but I take it no philosophical ice is cut by such examples." Here, Searle calls upon the philosophical concept of intentionality, which Brentano defines as "the hallmark of the Mental" and which can be associated with the set of faculties and

dispositions mentioned in the previous definition of intelligence. Searle is willing to grant to computers only "derived intentionality," a metaphorical shadow of the intentionality possessed by the resourceful computer programmers who create the illusion of intelligence. The introduction of intentionality provides another important way to characterize the debate between Turing and Searle: Turing believes that intelligent behavior is a failsafe indicator of the presence of intentionality, while Searle argues that the observation of such behavior gives us no means to determine whether the intentionality in question is original, true intentionality or illusory, derived intentionality.

As stated previously, The Turing - Searle debate as it has been framed here bears a strong similarity to the intellectual debate over the Argument from Design which was carried on during the eighteenth and nineteenth centuries. Although the primary focus of this debate as it was formulated by its participants was the existence of the Deity, it can be reinterpreted in more religion-neutral terms without compromising the essential positions of its contributors as a debate over the existence of purpose in the universe. The strongest proponent of the Argument from Design was the nineteenth-century theologian William Paley, whose treatise *Natural Theology* serves as an expression of the argument in its purest form. Like Searle, Paley relies heavily on a thought experiment which he uses to call upon certain intuitions common to the human experience. He asks his readers to imagine crossing a heath and encountering a pocket watch lying on the ground, then to reflect upon the probable cause of the watch's presence. In doing so, Paley invokes an overwhelming intuitive pull which forces any reasonable person to conclude that the only satisfactory explanation is the existence of an intelligent, purposeful watchmaker. From this point, the author extends the scope of this intuition to encompass the entire natural universe. As he writes, ". . . every indication of contrivance, every manifestation of design, which existed in the watch, exists in the works of nature; with the difference, on the side of nature, of being greater and more, and that in a degree which exceeds all computation. I mean that the contrivances of nature surpass the contrivances of art, in the complexity, subtlety, and curiosity of the mechanism; and still more, if possible, do they go beyond them in number and variety: yet in a multitude of cases, are not less evidently mechanical, not less evidently contrivances, not less evidently accommodated to their end, or suited to their office, than are the most perfect

productions of human ingenuity." Paley devotes most of the remainder of his text to the presentation of various observations from the natural world which indicate a level of complexity, design, and purpose far superior to that of human artifacts. In adopting this tactic, Paley uses the same basic operationalist tactic employed by Turing. He begins by noting that the concept of purpose as applied to human artifacts is characterized by certain observable traits such as design efficiency and complexity; he then incorporates the operationalist assumption that the presence of these traits is both necessary and sufficient for the applicability of the concept. Thus, Paley concludes that both the presence or absence of purpose in Creation and the nature of this purpose can be discovered through the careful observation of accessible phenomena in the natural world. In this aspect, Paley and Turing share common philosophical ground.

If Paley is the counterpart of Turing, then David Hume takes the role of Searle in the argument on design. Although Hume's *Dialogues Concerning Natural Religion* was published a quarter of a century before *Natural Theology*, it serves as a direct attack on the essential Argument from Design which Paley espouses. The dialogue pits Cleanthes and Demea, who represent respectively the forces of reason and dogmatic Christian belief, against Philo, who disagrees with the Argument from Design as presented by the other two. Cleanthes utilizes the argument in much the same manner as Paley; he observes a correlation of like effects shared by designed artifacts and natural phenomena, and from this concludes that rational purpose, the force responsible for these effects in artifacts, must also be the cause at work in the case of Nature. Philo begins his refutation with the observation that the similarity between the two effects is tenuous and imperfect at best, and thus that the operationalist inference made by Cleanthes requires a broad stroke of the imagination to include the regularities of Creation within the scope of the characteristic symptoms of artificial purpose. Later in the dialogue, however, he develops an argument with a much more significant impact. As he points out, the observable world contains not one but two concepts of purpose, each of which bears its own set of related observables. Artifacts have purpose which is imparted to them by their designers (an argument which foreshadows Searle's "derived intentionality"), while plants and animals possess a purpose which seems to derive from their own self-organization. Thus, the operationalist must decide which set of

characteristic phenomena the universe truly possesses before an assessment can be made of the nature of universal purpose. From here, Philo attacks Cleanthes' choice on this issue by stating that

> ...the operation of one very small part of nature, to wit man, upon another very small part, to wit that inanimate matter lying within his reach, is the rule by which Cleanthes judges of the origin of the whole; and he measures objects, so widely disproportioned, by the same individual standard. But to waive all objections drawn from this topic; I affirm, that there are other parts of the universe (besides the machines of human invention) which bear still a greater resemblance to the fabric of the world, and which therefore afford a better conjecture concerning the universal origin of this system. These parts are animals and vegetables. The world plainly resembles more an animal or a vegetable, than it does a watch or a knitting-loom.

By developing this tactic, Hume (through Philo) both delivers a preemptive blow to Paley's work and brings the analogy to Turing and Searle full circle. Just as Turing uses operationalist assumptions to deduce the existence of intentionality in digital computers, Paley uses the same technique to infer the existence of purpose in the universe. And just as Searle argues that Turing's operationalism cannot distinguish between original intentionality and intentionality derived from the programmers of the computers, Hume argues that Paley and his ilk cannot differentiate between purpose derived from a Creator and original purpose contained within the organic structure of the universe itself.

At this point, of course, Darwin intervenes. One of the few truly earthshaking publications in the history of science, *The origin of Species* establishes a new intellectual framework from which all that has gone before it must be re-evaluated. With the adoption of Darwin's paradigm, the issues of purpose and operationalism as debated by Hume and Paley are swallowed by a dense cloud of ideas which borrow from both writers but which fail to entirely vindicate either. The great contribution of Darwin to the debate is the reformulation of purpose as a teleological concept: purpose acquires a definition only relative to a given environment and is defined solely in terms of selection value within that environment. Whatever succeeds in being selected for in a given environment possesses sufficient complexity and design efficiency to have purpose attributed to it within that environment. This reformulation collapses

Hume's two classes of purpose into a single notion: just as a species which is moved to a new environment may lose its survival potential and thus lose its right to be attributed purpose relative to that environment, an artifact which is given a new task may not be attributed design purpose relative to its new function. The example of a pocket watch pressed into service as a doorstop illustrates this well: although the watch's complexity and regularity may still provide reasons to attribute design to the watch, its failure to succeed in the role of doorstop precludes one from attributing design purpose to it. Likewise, Darwin does not answer the ultimately theological question of whether species are consciously designed or not; he merely provides a teleological framework for the ascription of purpose. Thus, Paley is in some sense vindicated by Darwin's recognition of a single universal principle of purpose which can be derived through the observation of natural and artificial phenomena. However, Darwin's teleological formulation does not correspond exactly to Paley's concept of derived purpose. For Paley, purpose is derived from a supernatural Creator; for Darwin, however, if purpose is derived at all, it is derived from the complicated interrelation between the species and the environment. In the Darwinian world, purpose is no longer a purely metaphysical property which is unambiguously possessed by certain objects and which manifests itself through observable phenomena; instead, it is an epistemic notion which can be attributed to species only relative to a given environment and to the species' performance within that environment. Thus, Darwin refocuses the question of purpose from "'What *possesses* purpose?" to "In what contexts and under what circumstances can we *attribute* purpose?"

It is this astonishingly successful strategy which inspires Daniel Dennett to seek a position which both reconciles and overthrows Turing and Searle. Dennett's twenty-year commitment to the pursuit of a coherent notion of intentionality begins with his 1969 book *Content and Consciousness*; however, his ideas find their first clear expression in the 1971 publication of "Intentional Systems." Here, Dennett introduces the idea of stance adoption, the utilization of a certain attitude toward a certain seat of behavior as a method of predicating or describing the behavior in question. He first discusses the design stance, which can be viewed as an elaboration of Darwin's reformulation of purpose as already discussed. Dennett describes the various versions of the design stance as "alike in relying on the notion of *function*, which is purpose-relative or teleological." When animals or artifacts are analyzed from the design stance,

they are ascribed a purpose appropriate to their environment and then assumed to possess a design structure appropriate to that purpose. In the study of electronic devices, the design stance manifests itself through "black box" analysis; in the study of biological organisms, it appears as the adaptionist school of thought, a version of which Dennett defends in his paper on "Intentional Systems in Cognitive Ethology." From this Darwinian framework, however, Dennett abstracts to a higher level of stance adoption. Noting that the design stance becomes largely inappropriate when applied to the behavior of complex systems such as humans, animals and computer programs, he introduces what he calls the intentional stance. Adoption of this stance entails the assumption not only of environment-relative purpose but of purpose-relative rationality; by adopting the intentional stance, we assume that the systems under description have beliefs and desires appropriate to their environments and purposes and then predict their behavior by presuming that they will act rationally upon these beliefs and desires.

Dennett's reformulation of intentionality in these terms forces a wholesale reconsideration of the presumptions which Turning and Searle share in their debate. Like Darwin's concept of purpose, Dennett's definition of intentionality is teleological: it establishes a basic assumption of rationality and then justifies the attribution of the concept to any being whose behavior meets the terms of the assumption. And like Darwin's approach, Dennett's is a stance-relative concept: intentionality is no longer a property which can be possessed by a system, but one which can only be attributed to a system. This conception seems to justify Turing's essential vision in every particular. According to Turing, any computer program which is able to pass the Turing test can obviously be described through adoption of the intentional stance as well as its human competitors can; thus, by Dennett's definition, the computer is an intentional system as surely as the human mind is. There exists one substantial difference between Turing and Dennett, however, which proves to be fatal to the philosophical spirit, if not the letter, of the Turing test. Turing views intentionality as a metaphysical property which can be identified by the presence of certain observable phenomena; however, he does not define the property as simply the conjunction of the observables. He agrees with Searle that intentionality has an intrinsically phenomenological and unobservable component; however, he argues that the presence of intentionality's observable properties entails the presence of the metaphysical component as well. Dennett,

on the other hand, removes this component completely from his formulation of intentionality. As he writes, "We do quite successfully treat these computers as intentional systems, and we do this independently of any considerations about what substance they are composed of, their origin, their position or lack of position in the community of moral agents, their consciousness or self-consciousness, or the determinacy of indeterminacy of their operations. The decision to adopt the strategy is pragmatic, and is not intrinsically right or wrong . . . it is much easier to decide whether a machine can be an intentional system than it is to decide whether a machine can really think, or be conscious, or morally responsible." Thus Searle can claim some measure of vindication from Dennett's teleological position as well. Even though a machine which can pass the Turing test is by definition an appropriate target for adoption of the intentional stance, the act of adoption cannot confirm or deny the presence of those unique phenomenological properties and "causal powers" which Searle views as necessary for the existence of what he refers to as original intentionality.

In the end, however, neither Searle's nor Turing's conceptions of intentionality survive the transition to the teleological stance adoption of Darwin and Dennett. While Turing recognizes only a single form of primary intentionality and Searle divides intentionality into two types, original and derived, Dennett restricts the concept to a single notion which falls much closer to Searle's derived intentionality than to his idea of true intentionality. In his 1987 paper on "Evolution, Error, and Intentionality," Dennett argues persuasively to this point. Just as our artifacts derive their purpose from the environment in which we use them, they derive their intentionality in the same way. From coin-operated vending machines which "perceive" and "judge" quarters and slugs to sophisticated chess computers which "invent" plans and "pursue" goals, the products of design can have the intentional stance attributed to them only as a consequence of their performance within a given functional environment. Human beings, however, are nothing more than another species designed by the forces of evolution. Thus, just as Darwin demonstrates the purpose relativity of species, Dennett argues that the intentionality of all species, humanity included, is ultimately derived from the only truly primary level of intentionality: the level of evolutionary selection. We are machines designed by millennia of natural selection, and the fact that we can be described and predicated in terms of beliefs, desires, plans, and goals can be attributed entirely

to the motivating forces responsible for our presence and success within our environmental niche. This argument, then, brings the connection between Darwin and Dennett full circle. The arguments over operationalism advanced by Hume, Paley, Turing, and Searle are all swept aside by this powerful intellectual synthesis. The concepts of design, purpose, and intentionality are powerful ones which can afford us considerable explainatory and predicative power over a wide range of complex systems, from the myriad species of the natural world to the humans whom we interact with daily to the electronic machines which are rapidly changing the face of modern civialization. Ultimately, however, all of these attributions of purpose and intentionality must be relegated to the status of mere metaphors, of shadows derived from the single overriding purpose of natural selection. Through the blind manipulation of random genetic factors, an operation which seems wholly antithetical to the very essence of purpose and intelligence, nature has created a cosmic process endowed with a level of intentionality that dwarfs our own notion of intentionality and yet which provides the foundation for our use of the same notion. Dennett devises an intriguing passage which perfectly captures the difference between the position he shares with Darwin and the doctrine implicitly subscribed to by both Turing and Searle: "Aristotle said that God is the Unmoved Mover, and this doctrine announces that we are the Unmeant Meaners. . . [but] we are artifacts, in effect, designed over the eons as survival machines for genes that cannot act swiftly and informedly in their own interests. Our interests as we conceive them and the interests of our genes may well diverge — even though were it not for our gene's interests, we would not exist: their preservation is our original *raison d'etre*, even if we can learn to ignore that goal and devise our own *summum bonum*, thanks to the intelligence our genes have installed in us. So our intentionality is derived from the intentionality of our "selfish" genes! *They* are the Unmeant Meaners, not us!" This world-view of Darwin and Dennett is at once chilling and exhilarating. The choice is up to us: we can regard their ideas as a threat to the primacy of human authority, or as a challenge to press on in the pursuit of a comprehensive scientific vision of the forces which are ultimately responsible for our purpose, our intentionality, and our very existence.