

Reviews

The Turing Test: The Elusive Standard of Artificial Intelligence

JAMES H. MOOR (Ed.)

Dordrecht: Kluwer Academic Publishers, 2003

288 pages, ISBN: 1402012047 (hbk); \$114.00

In “Computing Machinery and Intelligence” (1950) Turing described an “imitation game” played by three people, an interrogator and two interviewees, one male (*A*) and one female (*B*). The interrogator communicates with *A* and *B* from a separate room (nowadays probably by means of a keyboard and screen); apart from this the three participants have no contact with each other. The interrogator’s task is to find out, by asking questions, which of *A* and *B* is the man. *A*’s aim is that the interrogator make the wrong identification. (Turing said, “The object of the game for the third player (*B*) is to help the interrogator. The best strategy for her is probably to give truthful answers.” (1950, p. 434).) In a second version of the game, a computer takes the part of *A* and a male or female (see below) the part of *B*. Now the interrogator’s task is to discover which of *A* and *B* is the computer; to do so he or she is permitted to ask any question, on any topic. The computer is allowed to do everything possible to force a wrong identification. Having described the computer-imitates-human game, Turing remarked that the question “Can machines think?” is “too meaningless to deserve discussion” and proposed replacing it by the question “Are there imaginable digital computers which would do well in the imitation game?” (1950, p. 442). Famously, he predicted that:

in about fifty years’ time it will be possible to programme computers . . . to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. (1950, p. 442)

The origin of Moor’s *The Turing Test* is a conference, *The Future of the Turing Test: The Next Fifty Years*, organized by Moor and held in 2000 at Dartmouth College. (The field of artificial intelligence (AI) is typically traced back to a 1956 Dartmouth conference, John McCarthy’s *Dartmouth Summer Research Project on Artificial Intelligence*. There Newell, Simon, and Shaw announced the *Logic Theorist*, designed to prove theorems from Whitehead and Russell’s *Principia Mathematica* and often assumed to be the first AI program.) The 2000 Loebner Prize Contest—the annual competition set up in 1991 to award a grand prize of \$100,000 to the first program to pass (a version of) the Turing test—was held alongside Moor’s conference. In this competition the test is not a three-player game. Instead, there are

several judges and several contestants, some human and some machine (the judges do not know the ratio of humans to machines); each judge interrogates every contestant in one-to-one interviews. Moor reports that in the 2000 competition “[n]o computer was mistaken for a human though in a few cases humans were judged to be computers!...the judges were 91% correct after five minutes of questioning” (p. 204). To date the (grand) Loebner Prize has not been awarded.

This collection has considerable depth and range (here I can discuss only a few contributions). It includes insightful papers on current debates by several of the foremost Turing scholars. The volume addresses the history, interpretation, criticism, and defense of the test, and alternative standards of intelligence in AI. It is genuinely interdisciplinary: the authors are philosophers, computer scientists, cognitive scientists, and social scientists, and the book also contains abridged transcripts of Loebner Contest interviews and examples of programming code. The discussions are accessible to students and much of the material is fascinating. Moor’s skilful editorship has produced a splendidly informative work. (Unfortunately the typesetting is poor.)

Copeland’s historical discussion of the state of artificial intelligence at the time of Turing’s “Computing Machinery and Intelligence” disproves the common belief that AI originated in the mid-1950s in the US. Turing was the first to carry out substantial research in the field, thinking about machine intelligence (the term in use in Britain, predating “artificial intelligence”) at least as early as 1941, and the first AI programs ran in 1951–1952 in the UK, at Manchester and Cambridge (where the world’s first electronic stored-program computers were built). Turing’s Bombe, which broke the German Enigma code during World War II, generated hypotheses as to the settings of the Enigma machine for the day’s message traffic, as a modern computer generates potential solutions to a problem (Copeland, 2004). In the early 1950s—at Manchester, where he wrote “Computing Machinery and Intelligence”—Turing even worked on what is now called “artificial life”.

On the interpretation of Turing’s test that has been orthodox for decades, the test yields an operational definition of thinking: a machine is intelligent (or thinks) if and only if its behavior in the imitation game matches that of a human being. However, Moor has already pointed out (1976) that in “Computing Machinery and Intelligence” Turing did not present the imitation game as a *definition* of (i.e., logically necessary and sufficient conditions for) intelligence or thinking. Copeland includes remarks by Turing in a 1952 radio broadcast (and elsewhere) which falsify the orthodox interpretation:

I don’t want to give a definition of thinking, but if I had to I should probably be unable to say anything more about it than that it was a sort of buzzing that went on inside my head. But I don’t really see that we need to agree on a definition at all. (p. 6)

If the orthodox interpretation is mistaken, what is the philosophical justification of Turing’s test? Moor states that a “plausible interpretation of the imitation game is to regard it as an inductive test... If a machine passed a rigorous Turing test,

then we... would have sufficient good evidence to infer that the machine was intelligent” (p. 202). This interpretation has the advantage that outcomes of the test are defeasible. The other advantages Moor claims are not so persuasive. He points out, first, that on the inductive interpretation we need not say that a device such as Block’s Jukebox, which passes the test only by utilizing a complex look-up table, is intelligent. In fact Turing provided a very different reply to the challenge (to his test) from such hypothetical devices. As Copeland notes, remarks in the 1952 broadcast suggest the following response: given the practical limitations on storage capacity, we could not build a brain-simulator that works by means of a look-up table—and even if we could, such a machine would take many years rather than minutes to answer the interrogator’s questions. Moor states, second, that on his interpretation the Turing test provides a scientific method of gathering evidence for the existence of intelligence in machines. However, the inductive interpretation fails to do justice to Turing’s various comments about the *nature* of intelligence; on Moor’s analysis, all the test tells us about intelligence itself is that it is a property for which success in the imitation game provides an indication.

Several authors—most profitably Copeland and Piccinini—turn to Turing’s accounts of the test outside his 1950 paper in order to refute objections to the test or clarify interpretation. For example, some recent discussions distinguish the standard reading of the rules of the imitation game—according to which a machine takes the part of *A* and a human being (man or woman) the part of *B*—from a “literal” reading, according to which *A* is played by a machine and *B* by a woman. In this volume Sterrett calls the latter game “the original imitation game” (OIG) test; the machine’s aim (like the man’s) is that the interrogator identify it as a woman. Traiger argues that Turing endorsed this form of the test. Against this exegesis of Turing, Piccinini provides a sweeping defense of the standard reading. Copeland points out that, in his 1952 radio broadcast, Turing said “The idea of the test is that the machine has to try and pretend to be a man . . . and it will pass only if the pretence is reasonably convincing” and that, in a May 1951 radio lecture, Turing presented the point of the test simply as determining whether or not a computer can “imitate a brain” (p. 8).

Sterrett argues that, irrespective of Turing’s intentions, the OIG test is the superior test of intelligence in machines. In her view, this test, but not the standard Turing test, provides a basis for comparing a machine’s skill at impersonation with a human’s; consequently, it is less sensitive to variations in skill among different interrogators. However, a machine passes the standard Turing test if it does no worse in the computer-imitates-human game than a man in the man-imitates-woman game (Copeland, p. 9). Hence the standard Turing test already possesses whatever advantage may result from a comparison of a machine’s and a human’s attempts at impersonation. Sterrett also claims that passing the OIG test depends upon intellectual resourcefulness rather than behavioral similarity, but Moor argues that the OIG test is the weaker of the two.

Moor claims that “Turing’s famous prediction . . . is disconfirmed by the results of the Loebner 2000 contest and the absence of any serious Turing test competitors

from AI on the horizon” (p. 197). (This was not Turing’s only prediction, as the book makes clear.) However, this prediction concerned the test as presented in “Computing Machinery and Intelligence” and, as noted above, the Loebner Contest does not follow that design—rather, as Copeland points out, it is closer to Turing’s 1952 version of the test. Copeland (and Moor himself) notes that the structure of the Loebner competition allows a bias against the machine: judges, it seems, are determined not to mistake a machine for a human being. Indeed, Turing said that judges in the 1952 version of the test might avoid error simply by “saying ‘It must be a machine’ every time without proper consideration” (Turing et al., 1952/2004, p. 495). Outcomes since Moor’s conference demonstrate this bias: in the 2003 competition, for example, on no occasion was a machine judged “definitely a human” (and only once “probably a human”) but on four occasions a human was judged “definitely a machine” (for more results, see the Loebner Prize Contest 2003 website). This bias is encouraged by the fact that, although in the 1952 formulation (as Copeland points out) Turing specifically excluded judges who were computer scientists, some judges in recent contests are plainly aware of typical AI programming strategies and weaknesses. (Zdenek’s paper illustrates the general hostility and suspiciousness of judges.)

Bringsjord proposes, as an alternative to the Turing test, the *Lovelace test* (in his 1950 paper Turing quoted Lady Lovelace, Charles Babbage’s collaborator, as saying that Babbage’s Analytical Engine did not “originate” anything). An artificial agent *A* passes this test if and only if *A*’s designer (or someone with equivalent knowledge and resources) cannot explain *A*’s output by appeal to *A*’s architecture, knowledge-base, and core functions. (*A*’s output is not the result of a fluke hardware error.) Bringsjord seems unaware that (as Copeland has shown) Turing too linked intelligence with explicability: Turing said “If we are able to explain and predict [a device’s] behaviour . . . we have little temptation to imagine intelligence” and “one might be tempted to define thinking as consisting of ‘those mental processes that we don’t understand’” (Copeland, 2004, p. 491). Bringsjord points to the machine candidates in the Loebner Contest as evidence that Turing’s test “is such as to cultivate tricksters” (p. 215). In fact, the outcomes of these competitions suggest that a device whose behavior can be easily explained by a judge familiar with AI programming tricks is unlikely to pass Turing’s test.

Bringsjord fails to provide any argument that passing the Lovelace test is either necessary or sufficient for, as he puts it, “the presence, in computers, of such ‘deep’ phenomena as thought and consciousness” (p. 215). The Lovelace test is remarkably tough. Bringsjord says that passing it “may require a kind of causation beyond the bounds of ordinary causation” (p. 237) and that such “agent causation” may be the kind of autonomy possessed by human beings. However, he offers no reason to think that humans would pass this test or that it captures what we mean by “originality”. Harnad proposes another exacting test, the *Robotic Turing test*: in this the machine candidate “must be an embodied robot, capable of nonsymbolic, sensorimotor interactions with the world in addition to symbolic ones, all Turing indistinguishable from our own” (p. 266). The advantage of this test is that it evades Searle’s (1980)

Chinese room argument, since the latter targets only “an implementation-independent implementation of a formal symbol system” (pp. 262–263). However, the Chinese room argument is much debated and Harnad has joked that he and Searle are perhaps the only people on earth who think it valid.

Turing’s test emerges from this volume undefeated by celebrated objections and of continuing philosophical importance.

References

- Copeland, B. J. (Ed.). (2004). *The essential Turing*. New York: Oxford University Press.
- Loebner Prize Contest 2003: The Contest*. (n.d.). Retrieved from <http://www.surrey.ac.uk/dwrc/loebner/results.html>
- Moor, J. H. (1976). An analysis of the Turing test. *Philosophical Studies*, 30, 249–257.
- Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3, 417–424.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59, 443–460.
- Turing, A. M., Braithwaite, R., Jefferson, G., & Newman, M. (2004). Can automatic calculating machines be said to think? In B. J. Copeland (Ed.), *The essential Turing* (pp. 494–506). New York: Oxford University Press. (Original work published 1952).

DIANE PROUDFOOT

Department of Philosophy

University of Canterbury

Christchurch, New Zealand

Email: Diane.Proudfoot@canterbury.ac.nz

Reflections and replies: Essays on the philosophy of Tyler Burge

MARTIN HAHN & BJØRN RAMBERG (Eds.)

Cambridge, MA: MIT Press, 2003

424 pages, ISBN 0262582228 (pbk); \$45.00

Meaning, basic self-knowledge, and mind: Essays on Tyler Burge

MARÍA J. FRÁPOLLI & ESTHER ROMERO (Eds.)

Stanford, CA: CSLI Publications, 2003

299 pages, ISBN 1575863464 (pbk); \$25.00

Whether you suspect your bottled water has been switched for potable XYZ, or fear your arthritis has spread from knee to thigh, reflections on the physical and social constraints on the meaning of our words, and the individuation of mental states themselves, now play a central role in contemporary philosophy of language and mind. The semantic content of proper names, natural kind terms (and even common nouns like ‘sofa’), and representational states has generally come to be viewed as governed by causal factors “outside the head” rather than merely by internal mechanisms acting as the cognitive linkage relating the mind and its thoughts to the world and its objects. Tyler Burge stands at the center of these developments.

On one hand, Fregean inspired accounts of meaning exploit the semantic discrepancy between different but co-referring names. By contrast, externalist accounts of meaning and anti-individualist accounts of mind highlight the semantic discrepancy among referentially distinct but homophonous words: ‘Phosphorus’ and ‘Hesperus’ are two different types of words with one and the *same* referent, but ‘water’ (as spoken on Earth) and ‘water’ (as spoken on Twin Earth)—like ‘arthritis’ (as spoken in English) and ‘arthritis’ (as spoken in Twin English)—appear to be the same type of word with *distinct* referents (despite how they might be incorrectly used by some). And while traditional sense theories generalize from scenarios where there is a many-one relationship between distinct names (e.g., ‘Phosphorus’ and ‘Hesperus’) and a single referent (Venus), Putnam’s semantic externalism and Burge’s anti-individualism generalize from the inverse—scenarios where there is a many-one relationship between distinct referents (e.g., H₂O and XYZ) and what appears to be a single type of word (‘water’), or (as with Burge) a many-one relationship between distinct mental states and physically identical Doppelgangers. For however phonetically identical, ‘water’ and ‘water’ are *not* the same type of (ambiguous or even indexical, see below) word, and Burgean twins are *not* in the same mental state. Both the individuation of words and the individuation of psychological states are, if externalism and anti-individualism are correct, conditioned by environmental, social and historical factors. Yet there are differences between Putnam and Burge. Putnam’s (1975) arguments feature the actual physical environment’s role in determining what our words mean, while Burge has also detailed the role our sociolinguistic environment, and in particular, our linguistic allegiances in the form of semantic deference, play in determining not only what our words mean, but what kinds of thoughts we can entertain. Unlike Putnam’s twin cases, it would seem that the nature of our sociolinguistic conventions—our dictionaries—also play a role in specifying the content of thought.

There is, however, another difference between Putnam and Burge. Nothing comparable to the publication of Pessin and Goldberg’s (1996) *The Twin Earth Chronicles* has been prepared specifically on behalf of Burgean scholarship issuing from his “Twin English” thought experiments. The two new volumes under review are now correcting for this discrepancy. (Another volume—*Themes from Burge*—also forthcoming from CSLI, further amplifies this crescendo in Burgean philosophy.) Together, these volumes may very well provide an insider’s preview of what *The Twin English Chronicles* might eventually look like: a series of articles dispatched from a place where our psychological states are at least partially conditioned by what our words conventionally refer to, and what those states are normally supposed to represent.

Both *Reflections and Replies* (hereafter, *R&R*) and *Meaning, Basic Self-Knowledge, and Mind* (*MBS*) are the product of international conferences devoted to Burge’s work. The papers included in *R&R* were originally presented at a 1993 conference held at Simon Fraser University. Indeed, the table of contents for this *festschrift* reads like a virtual seating chart for a one-of-a-kind banquet table of philosophers of mind and language: Ned Block, Noam Chomsky, Keith Donnellan, Fred Dretske, James

Higginbotham, Brian Loar, Joseph Owens, Christopher Peacocke and Barry Stroud, among others. *MBS* is a collection of papers delivered at a conference entitled “The First-Person, Other Minds, and Interlocution” sponsored by the Philosophy Department at the University of Granada in 1996. While most of the papers in this volume were prepared by philosophers associated with the University of Granada (or other universities in Spain), the editors also invited a number of non-European scholars to contribute to the volume. Articles by Christopher Gauker, Steven Davis, and Martin Davies were also included. Both volumes, however, retain the feel of a conference dialogue, as Burge responds to each paper individually. In fact, Burge’s replies can be as extensive as they are intensive: *MBS* includes over fifty pages of Burge’s response—making it the single largest contribution; *R&R* accommodates almost two hundred pages of Burge’s reactions—a book unto itself.

Some of the papers are more narrowly focused on particular issues or even specific debates among scholars writing within the broad philosophical arena of “externalism” and “anti-individualism.” In *R&R*, Calvin Normore re-reads Descartes’ *Meditations* with an eye to the issues raised by the externalist orientation of Burge’s philosophy of mind, and Burge’s philosophy of perception (ch. 1). In the same volume, Martin Hahn’s “When Swampmen Get Arthritis” diagnoses superficial similarities to reveal deep discrepancies between Davidson’s holistic and interpretationist externalism and Burge’s anti-individualist version of externalism (ch. 3), while Ned Block’s “Mental Paint” article defends the non-representational nature of phenomenal qualia (ch. 9). The papers in *MBS* develop several debates, with some contributors directly challenging Burge, others comparing Putnam’s “physical” externalism with Burge’s “social” externalism, and still others attempting to evaluate disputes more generally. In *MSB*, for instance, Gauker (ch. 1) criticizes Burge for circularly relying on “expressivist” assumptions which undermine Burge’s own model of anti-individualist linguistic communication, while Marqueze (ch. 3) assesses how Gregory McCulloch’s (1995) presentation of externalism fares against the standing arguments against “orthodox externalism” developed by Bilgrami (1992). The discussion pursued by Tobies Grimaltós (ch. 2) explores the differences between Putnam’s Twin Earth and Burge’s Twin English by contrasting what he calls “referential” content with “deferential” content.

As might be expected, however, a few “big ticket” items receive more continuous treatment from author to author, as well as from book to book. One concerns the compatibility of Burge’s anti-individualism with introspective self-knowledge of one’s own thoughts and related issues of skepticism and immunity to error. (Two others are mental causation and indexicality, see below). Chapters 5–9 of *MSB* speak directly to this issue of reconciling anti-individualism with a privileged sense of *a priori* knowledge concerning our own thoughts. By contrast, the papers which speak to the possibility of first-person knowledge of externally individuated mental content do so in varying degrees, and are scattered throughout *R&R*. These are Stroud’s “Anti-Individualism and Scepticism” (ch. 2), Peacocke’s “Implicit Conceptions, Understanding, and Rationality” (ch. 7), Bernard Kobes’ “Mental

Content and Hot Self-Knowledge” (ch. 10), and Loar’s “Phenomenal Intentionality as the Basis of Mental Content (ch. 11).

In response to Kobes’ article, for instance, Burge returns to the original problem:

In my (1988), I asked why the fact that we have only empirical access to causal relations that fix the nature of our thoughts does not entail that we cannot know that we are thinking such and such unless we engage in empirical investigation that shows the conditions for thinking such and such are satisfied. [I said that the answer] can be seen as a series of variations on the point that one must start somewhere. (p. 425)

In particular, while Burge concedes that the external conditions for thinking a particular thought must, as a matter of causal necessity, be presupposed in the thinking, one need not know what these external conditions actually are in order to think the thought. That the external conditions actually obtain is enough; knowledge of these conditions is not necessary. The sort of environmental dependencies implied by externalism are not an affront to basic self-knowledge simply because these external enabling conditions need not be known. They must only be in effect. But as Davies points out in “Externalism, Self-Knowledge and Transmission of Warrant” (*MBS*, ch. 5), it is as if “Burge is allowing that in thinking that water is wet, or in thinking that I am thinking that water is wet, I presuppose or assume that the conditions necessary for me to think that thought do obtain. In that case, my knowledge that I am thinking that water is wet is not knowledge that I can ‘have without making any assumptions about the external physical world’” (p. 118). In response, Burge again emphasizes the distinction between conditions which enable a thought, and knowledge of those conditions: “A child can think that water is wet,” says Burge, “without having the concepts *condition, environment, causal relation between environment and individual subject, normal*, and so on . . . It is an impersonal relation between the thinking and actual principles or conditions governing its possibility” (*MSB*, p. 264).

The problem with this response is that this automatic inclusion relationship is liable to be too *impersonal*—so impersonal that even if you *are* thinking about water, there is no way for *you* to know that you are thinking about water. Burge’s compatibilist solution is, as it were, a Pyrrhic victory of sorts. Several authors in *MSB* share this suspicion. As Carlos Moya notes in “Externalism, Inclusion and Knowledge of Content” (*MSB*, ch. 8), “Some philosophers feel that the inclusion model makes self-knowledge into a rather anaemic cognitive achievement, so much so that its entitlement to the dignity of knowledge could be justifiably questioned” (p. 170). (In response, Burge takes issue with Moya’s claim that his compatibilism “draws too heavily on an externalist, reliabilist view of justification” (p. 279).) According to Burge, when you are thinking about water, then you can be sure through cogito-like self-reflective introspection that you are indeed thinking about water. But what about those who are not sure whether they are thinking about water—as opposed to, say, XYZ *twater*—in the first place? What about them? And are we not all in this position without empirical investigation? The problem,

then, is not that there is potential dislocation between first-order and second-order thoughts, rather just the opposite: second-order thoughts will directly inherit all the clarity, or as the case may be, unclarity of our first-order thoughts.

So while it might be the case that we are automatically guaranteed basic self-knowledge of our thoughts—whatever they may be—we seem to be in no *a priori* position to determine what they actually are. It seems not to be sufficient, argues Daniel Quesada in “Basic Self-Knowledge and Externalism” (*MSB*, ch. 9) to maintain that:

If background conditions are different enough so that I am thinking different thoughts, they will be different enough so that the objects of reference and self-ascription will also be different. So no matter how my thoughts are affected, no matter how I am switched around, I will be correct in self-ascriptions of content that are correctly expressed in cogito-that-clause form. (p. 190)

Maybe knowledge of content does not require a “transparency of content” principle (Boghossian, 1994) sufficient to distinguish *water* thoughts from *twater* thoughts. This much Burge is quick to press, for he considers it to be extremely misleading to count (these switching cases) as the subject’s ‘lack of self-knowledge of his own contents’:

I have discussed at some length the difference between ordinary knowledge of one’s contents in ordinary self-attributions that count as knowledge of one’s mental states and events, on the one hand, and an ability to explicate and individuate one’s concepts in a global way (so as to deal with all contingencies), on the other. I believe that individuals have no special first-person authority in the latter case. (*MSB*, p. 256)

In anticipation of this sort of move, María Frápolli and Esther Romero complain in their article “Anti-Individualism and Basic Self-Knowledge” (*MSB*, ch. 6) that this minimalist sanctuary for basic self-knowledge “diminish[s] the first person authority to an almost trivial role” (p. 144). Similarly, Antoni Benejam (*MSB*, ch. 7), like Quesada (*MSB*, ch. 9), objects that the Cartesian ideal of self-reflective clarity and distinctness loses its methodological force once we are no longer positioned to “get at what is essential in the content, that without which it cannot be conceived, that which necessarily constitutes its essence, its ‘true and eternal nature’” (p. 153). “[W]e should not be blamed if a feeling of insatisfaction remains,” says Quesada, for “According to what it has been told so far, it would appear that there is no conflict whatsoever between basic self-knowledge and externalism, only because the self-ascribing beliefs that constitute such knowledge are, so to speak, immunized against changes in the environment” (p. 192).

A second issue no more avoidable than basic self-knowledge concerns the causal efficacy and explanatory utility of “folk psychological” appeals to wide content given the local “in-the-head” supervenience of the mental on the physical. Two chapters in *R&R* (chs. 8 and 12) and the last two chapters in *MSB* (chs. 10 and 11) discuss the explanatory value and causal force of anti-individualistic content—what one contributor refers to as the action-at-a-distance threat of “crazy causality”

(“Individualism, Internalism, and Wide Supervenience,” ch. 11) and what Dretske (*R&R*, ch. 8) identifies as his “epiphobic” fear—the threat of mental epiphenomenalism. But whereas Dretske presses Burge to explain why he is not also troubled by the superfluous causal powers of wide mental content, in his contribution to *R&R*, Chomsky (ch. 12) dismisses the cottage industry of Twin Earth/Twin English “folk semantics” as unreliable speculation and largely irrelevant. For Chomsky, the sciences of mind, like his model for “I-linguistics,” is internalist and individualist. Burge’s responses to these complications, worries, and assertions is simple: token causal potency should not be confused with taxonomic typing. “Internalist” psychology may be all that is needed to explain *how* an organism represents (or how a vending machine works), but determining *what* it represents (or what such a mechanical device is *for*) requires more. “Anti-individualism is,” as Burge reminds us in *MSB*, “a view about the nature or individuation of mental states” (p. 289). In fact, individualists have it wrong because “Causal power individuation is not . . . prior to psychological state individuation in psychology” (p. 256).

A third, but less clearly visible, topic concerns the relationship among names, natural kind terms and indexical expressions, and how indexical expressions interact with Twin Earth and Twin English thought experiments. In particular, Burge has been fighting a rear guard action against indexically inspired counter-interpretations of twin scenarios. So while Putnam (1975) originally expressed some sympathy for interpreting ‘water’ (as said on Earth) and ‘water’ (as said on Twin Earth) as expressions with a (“hidden”) indexical component contextually pointing toward H₂O on Earth and XYZ on Twin Earth, Burge is opposed to this indexical strategy. (And in his responses, Burge does not overlook any opportunities to emphasize that even Putnam has disowned this indexical interpretation of twin cases.) In his contribution to *R&R* for instance, Donnellan proposes an indexical rule reformulation of Putnam’s Twin Earth which preserves the twin’s psychological equivalence, and then discusses its relationship to Burgean cases of Twin English (ch. 4). But even Burge’s response to Donnellan is not the final word. In “Anti-Individualism, Indexicality, and Character” (*R&R*, ch. 5), Owens turns the table on indexical approaches by arguing that the meaning of indexical expressions is itself determined anti-individualistically, while in *MSB*, Davis explicitly defends Burge’s anti-indexical account against Donnellan’s advances (ch. 4).

What *R&R* offers in terms of established philosophical luminaries, *MSB* broadens in terms of a horizon for lesser known figures to also rise. The volumes also differ in editorial polish, precision and ultimately, purpose. On the one hand, proofreading errors in *MSB* and a certain non-native awkwardness among some of the contributors can lead to a somewhat bumpy read. And though both volumes include a combined subject-author index, the index in *R&R* is detailed and accurate while the index in *MSB* is shorter on both. On the other hand, the editors of *R&R* have made a real effort to guarantee their volume scholarly utility. Not only does *R&R* provide an up-to-date bibliography of Burge’s work, it also includes a comprehensive listing of secondary sources on Burge. This reference list is key to understanding the impact Burge has had on contemporary philosophy of mind

and language, and is an invaluable research tool. Its 19 pages make for a “one-stop shopping” list of recommended reading before one can make the journey to Twin Earth or begin to understand Twin English.

References

- Bilgrami, A. (1992). *Belief and meaning*. Oxford, England: Blackwell.
- Boghossian, P. (1994). The transparency of mental content. In J. Tomberlin (Ed.), *Philosophical perspectives: Vol. 8. Logic and language* (pp. 33–50). Atascadero, CA: Ridgeview Press.
- Burge, T. (1988). Individualism and self-knowledge. *Journal of Philosophy*, 85, 649–663.
- Burge, T. (1996). Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society*, 96, 91–116.
- McCulloch, G. (1995). *The mind and its world*. London: Routledge.
- Pessin, A., & Goldberg, S. (Eds.) (1996). *Twenty years of reflection on Hilary Putnam’s “The meaning of ‘meaning’.”* Armonk, NY: M. E. Sharpe.
- Putnam, H. (1975). The meaning of ‘meaning’. In H. Putnam, *Philosophical papers: Vol. 2. Mind, language and reality* (pp. 215–271). Cambridge, England: Cambridge University Press.

REESE M. HEITNER
Department of Philosophy
The Graduate Center
City University of New York, USA
Email: reeseheitner@hotmail.com

Taking Action

SCOTT H. JOHNSON-FREY (Ed.)
Cambridge, MA: MIT Press, 2003
413 pages, ISBN: 0262100975 (hbk); \$60.00

The 13 new essays collected in this volume are united by a common persuasion—that neurologists, neuroscientists, and psychologists often make the mistake of crudely partitioning the terrain of brain functions into three isolated domains: perceptual, cognitive, and motor. The dissenting view shared by all of the contributors to this volume is that these three domains are far more intimately connected than they are traditionally taken to be. More specifically, it is argued that intentional actions, in particular, involve *complex interactions* between them.

The essays fall under five sections: “Perception and Action,” “Intention and Simulation,” “Gesture and Tool Use,” “Sequencing, Coordination, and Control,” and “Learning and Movement.” In each case, the general aim is to show how the best (indeed sometimes the only) way to understand the related mechanisms of behavior is to start thinking across the “artificial boundaries” created by the scheme of partitioning the brain into three principle domains. All of the arguments are supported by experiments, the details of which are laid out in a plain yet meticulous

fashion. Be that as it may, however, the conclusions reached occasionally rely on unpersuasive inferences, typically encouraged by conceptual confusion.

In the first section, "Perception and Action," A. David Milner and Richard T. Dyde describe different experiments that make trouble for traditional arguments from illusory perception, whose aim was to show that there was no interaction between ventral and dorsal systems. Similarly, James A. Danckert and Melvyn A. Goodale appeal to differential representations of upper and lower visual fields to demonstrate that a process within one system can behave very differently in the context of processes within another and that, consequently, it would be perilous to study perception, cognition, and motor processes in isolation.

The second section, "Intention and Simulation," opens with Yves Rossetti and Laure Pisella's account of how evidence from tasks that involve inhibiting direct responses to specific stimuli supports the view that the difference between automatic and intentional action is one of *degree* rather than of *kind*. If they are right about whether the evidence supports this view, then it cannot be the case (as many believe) that the dorsal stream is involved only in automatic actions and the ventral only in intentional ones. But, failing to distinguish between actions that are *intentional* and actions that are *voluntary* (pp. 90–91), Rossetti and Pisella unfortunately appeal to experiments which suggest that an action is voluntary to support the conclusion that intention is in the offing (p. 98). Their additional failure to distinguish between *transitive* and *intransitive* consciousness leads them to also make the unguarded concluding remark that whenever we act intentionally we have conscious knowledge of our actions; yet, all that follows from their argument is that one cannot intend anything unless one has at some point been conscious (in the intransitive sense). After all, it is perfectly sound to attribute intentions to people who are asleep.

Next comes Marco Iacoboni's summary of the first results of a research program on the basic mechanisms underlying social behavior that is guided by the observation and imitation of other people's actions. Iacoboni rightly concludes from this research that it is more than likely that our witnessing and imitating the actions, intentions, and emotions of *others* triggers the same neural activity associated with *our own*. However, his further claim—that his finding that neurons in the ventral premotor cortex and in the anterior part of the posterior parietal cortex discharge both when a monkey *performs* an action directed toward an object, and when the monkey *observes* another individual performing a similar action, suggest that "an individual can recognize an action made [sic] by others because the observed action evokes a discharge in the same neurons that fire when the individual performs the action" (pp. 108–109, emphasis in the original)—is at best misleading. No doubt the fact that there is a discharge in the same neurons tells us much about the neural substructures of both the actions and the observations which they *enable*; but it simply does not follow, and is not true either, that it is *because of this* that we recognize certain actions performed by other people. The explanations of why we recognize them (e.g., because we recall similar actions, or have performed them ourselves) are successful *regardless* of what we discover about our neurology.

Unfortunately, Iacobani's misconceptions do not end here. Towards the end of his essay he briefly refers to "three main views" on the nature of intentionality: that it is a symbol-based system (Fodor), that it is nonlinguistic (Dennett), and that it is essentially normative, thus determined by social practices (Heidegger). So far, so good. But Iacobani continues:

We assumed that the patterns of brain activity we detected would lead us to discover how intentionality actually works. If language- or math-related areas were activated in the critical experimental comparisons, the first view would be correct and intentionality would probably be symbol based. If sensor motor and prefrontal areas were activated in the critical experimental comparisons, then the second view would be correct. Finally if the "social brain" (limbic system, orbit frontal cortex) and sensor motor areas were mostly activated in the critical experimental comparisons, then the third view would be correct. (p. 131)

This is to grossly misunderstand the very nature of what is being debated here. For if the view attributed to Heidegger—which is also the view Iacobani takes his experiments to prove right—is the correct one, then no amount of neurology could ever tell us anything about intentionality, for it is a central tenet of this view that all discerning criteria for intentionality lie within social and normative practices (which cannot be captured in purely physical terms). Likewise, if the Dennettians are right, then there is no such thing as a truly normative practice (for such a practice *requires* a linguistic understanding of the mental), and the "social brain" is only responsible for our adopting a normative *stance*. Finally, if Fodor's view is correct, then social practices, and normativity itself, would be symbol-based anyway, so even if the "social brain" was *permanently* activated during the experimental comparisons, it would still not follow that the "social" view of intentionality was correct since, as we have already seen, this view rejects the idea that normativity can be captured in purely physical terms; if this is true then, *inter alia*, normativity cannot be reduced to mere symbol-manipulation. Philosophical mistakes aside, Iacobani's paper paves the way for Marc Jeannerod's essay on action simulation, which provides a useful taxonomy according to which many brain regions believed to be exclusively in perception or motor control can be engaged by purely cognitive processes that involve no overt movements.

Section three is comprised of two essays, one on gesture and one on tool use. In the first of these, Angela Sirigu et al. draw on their studies of patients with manual apraxia (i.e., the inability to perform certain purposive actions, as a result of cerebral disorder) with the aim of showing that the parietal lobe for human hand gestures is modularly organized. To this end, the research conducted by Sirigu et al. includes neuropsychological evidence from patients having lesions in the parietal cortex. In the second essay, Scott H. Johnson-Frey (the volume's editor) collects disparate observations concerning the neural substrates of primate tool use and explains why these suggest that the processes and mechanisms that underlie tool use actions span multiple systems.

As with the previous section, each of the themes of section four, "Sequencing, Coordination, and Control," has a separate essay devoted to it. Thus, Richard B. Ivry

and Laura L. Helmuth (who focus on brain lesions) provide recent insights into how the brain flexibly *sequences* individual movements, Elizabeth A. Franz (who focuses on bimanual actions) into how it *coordinates* multiple limbs, and Michel Desmurget and Scott Grafton (who integrate a wide range of findings) into how it *controls* feed forward and feedback mechanisms. All three essays aim to illustrate how the processes in question cut across the domains traditionally assigned to them.

The fifth and final section, “Learning and Movement,” is concerned with so-called “internal representations” and the neural substrates in which these are realized. Thus, Camillo Padoa-Schioppa and Emillo Bizzi look into how the brain *enables* the acquisition of new actions, M. Rotte into how it *cope*s with the challenge of controlling action when compromised by disease, and Wolfgang Kruse et al. into the sensor-motor transformations which it must “*solve*” in order for us to catch moving targets. Needless to say, *interaction* is the word all-round.

The writing throughout the volume is extremely lucid, and the evidence provided by the numerous experiments conclusive. It is therefore a great shame that the success of such a ground-breaking volume is occasionally tainted with the infelicity of some of its contributors misinterpreting the findings of their own experiments. No doubt such confusions do not affect the volume’s impact at large, but it is nevertheless unfortunate to see earnest practitioners misunderstand the implications of their own experiments, thereby laying themselves and their cause open to criticism which might have otherwise been avoided. Moreover, these errors betray a worrying lack of appreciation of where the value of their own research truly lies: not in resolving philosophical dispute, but in making scientific advances with regard to how the brain functions.

CONSTANTINE SANDIS
Oxford Brookes University
Harcourt Hill Campus
Oxford OX2 9AT, United Kingdom
Email: csandis@brookes.ac.uk

Perceptual Dynamics: Theoretical Foundations and Philosophical Implications of Gestalt Psychology

FREDRIK SUNDQVIST

Göteborg: Acta Philosophica Gothoburgensia, 2003

248 pages, ISBN: 9173464864 (pbk); \$24.00

Gestalt psychology was founded around the 1920s by Wertheimer, Köhler, and Koffka. At the time, it presented a revolutionary approach to cognition that, to this very day, has continued to be controversial. Some reviews depict Gestalt psychology as something from the past: It may have given us nice demonstrations of perceptual grouping principles, but it lost its value to future research once Köhler’s physical

brain model had been disproved in the 1950s. Such funeral orations are countered by eulogies arguing that the founding fathers of Gestalt psychology were ahead of their time and erected the pillars on which much of modern cognitive science stands. Sundqvist's *Perceptual Dynamics* falls in the latter category.

The first part of the book discusses the original Gestaltist ideas that, in the second part, are argued to gain fresh relevance by recent developments in cognitive science. Indeed, for several decades, Gestalt psychology may have been at the periphery of cognitive science but, in recent years, it is a subject of renewed interest. There is a growing awareness that further elaboration of the original Gestaltist ideas is needed for modern developments to prosper. In view of this Gestalt revival, Sundqvist's book is timely. It does not refer extensively to specific recent models but, then again, its main objective is to discuss underlying theoretical and philosophical ideas.

The temporary move to the periphery occurred, as Sundqvist argues, partly because the Gestaltists felt unable to live and work in Nazi-Germany but neither found fertile soil for their ideas elsewhere, and partly because existing technologies were not advanced enough to either prove or disprove their ideas. As Sundqvist argues, the 1950s discussion about Köhler's physical brain model actually ended undecided, and the then rising technology of single-cell recording pushed cognitive science in another direction, but modern brain imaging technology may prove that Köhler was right after all. I have reservations about the current brain imaging hype (see Koenderink, 1999) but I agree that science is modulated by the zeitgeist which also enabled the rise of Gestalt psychology.

Nowadays, Gestalt psychology may be known primarily for its laws of perceptual grouping, but these laws were actually just demonstrations of fundamental ideas that had been triggered by developments, around 1900, in physics. These developments in physics cast strong doubts on then-existing approaches to cognition and called for a new approach that would take better account of the fact that the brain is, in the end, just another physical system. Physical systems tend to settle in stable states characterized by a maximum of regularity, symmetry, and simplicity. The Gestaltists therefore postulated that the brain, in reaction to stimuli, also tends to settle in such stable states.

This tendency they called the 'Principle of Prägnanz', and these stable brain states they called 'Physical Gestalts', in analogy to von Ehrenfels' notion of Gestalt qualities referring to conscious experiences evoked by stimuli. To specify the relationship between physical Gestalts and phenomenal Gestalts, they put forward the hypothesis of psychophysical isomorphism which, in Köhler's words, means that psychological facts and underlying events in the brain resemble each other in all their structural characteristics.

Sundqvist argues that these original Gestaltist ideas link up well with current ideas about the dynamics of the brain's neural network. More specifically, Sundqvist focuses on present-day approaches that start from either dynamic systems theory (DST) models or connectionism (network models).

DST is the theory that, around 1900, rose in physics and that, nowadays, is also known by the glossy but misleading name ‘chaos theory’. It uses differential equations to capture nonlinear (and seemingly chaotic) behavior of systems over time in deterministic descriptions: For any given system state, the differential equations specify unambiguously the next state. One of many DST applications is the description of electro-magnetic (EM) fields. Inspired by DST, Köhler hypothesized that electrical currents in the brain maintain an EM field that, in turn, may modulate activity in different parts of the brain (even if these parts have no direct neural connection). A physical Gestalt then results from this interaction between local activity at the neural level, on the one hand, and the global EM field, on the other hand. Sundqvist now argues that Köhler’s idea complies with recent findings about synchronicity: Neurons in different parts of the brain may fire synchronously, suggesting that these brain parts communicate non-neurally. Although the jury is still out on the role of EM fields, synchronicity seems indeed an implementation of the Gestaltist motto that the whole (here, a physical Gestalt) is different from the sum of the parts (here, local activity at the neural level).

Connectionism rose in the 1980s and reflects—according to Sundqvist—another implementation of this Gestaltist motto. Connectionist models, or network models, are also known by the glossy but misleading name ‘neural networks’. Networks in network models may have a structure resembling the neural structure of the brain (i.e., nodes and links between the nodes) but are in fact distributed data representations as used, since the 1950s, in computer science and studied in graph theory. An everyday example of a distributed representation is a road map in which routes between places are not displayed separately (as route planners do) but such that common parts are displayed effectively as common parts. Network models can, at best, be said to add network processes (e.g., DST-describable forms of activation spreading) that might resemble brain processes better than the usual network processes in computer science do. Nevertheless, I agree that networks form suitable tools to investigate cognition within the Gestaltist tradition.

In 19th-century psychophysics, the percept of a noisy signal was already known to be determined by the signal-to-noise ratio rather than by the signal or the noise as such. Some time later, the Gestaltists hypothesized that cognition works with “relations between things” rather than with “things” as earlier approaches supposed. As Sundqvist argues, this hypothesis was inspired by a DST-related trend in physics to describe the world no longer in terms of physical objects but rather in terms of interacting force fields. The Gestaltist hypothesis means that a whole we think we perceive (i.e., a phenomenal Gestalt) is the result of an interaction between the parts that the sensory system picks up from a stimulus. As Sundqvist argues, this idea became an integral part of Gibson’s ecological approach and network models too implement this idea. Indeed, these models thrive on networks that represent relations between parts that interact to arrive at wholes. Some network models represent wholes by

network nodes, but better in line with the Gestaltist ideas are network models that take wholes to be reflected by patterns of activation.

I think Sundqvist did a fine job by strengthening the position of Gestalt psychology in present-day cognitive science. Furthermore, his account of the impact of Gestalt psychology on philosophers like Wittgenstein was revealing to me. The book is therefore a timely and valuable contribution to the Gestalt revival that is taking place, but—as I suggest below—it might have given a much better embedding of representational approaches to cognition.

It is unfortunate that Sundqvist seems to have a blind spot for the potential value of representational approaches to cognition. Many current representational approaches fit in the Gestaltist tradition, particularly those that explicitly aim at capturing the concept of *Prägnanz* in terms of descriptive simplicity (for a review, see van der Helm, 2000). This is not recognized in the book. In fact, Sundqvist rejects representational approaches because they allegedly portray the brain as being merely a computer-like symbol manipulator. This objection is raised more often against representational approaches but it actually confounds models with what is being modeled. Representational models indeed use symbolic descriptions, but so do DST models and network models. Representational models use stimulus descriptions to model structural characteristics of cognitive states, just as DST models use differential equations to model transitions between brain states. Network models claim to do a bit of both, but most central is their usage of distributed representations to model interactions between stimulus features.

This may also show that the distinction between these three types of models is primarily a distinction between formal tools. Particular tools may be convenient to model particular aspects of cognition but should not be mistaken for the underlying ideas about cognition as a whole. Different tools may well complement each other within, say, the Gestaltist framework of psychophysical isomorphism that tries to relate stimuli to both physical Gestalts (i.e., brain states) and phenomenal Gestalts (i.e., cognitive states).

For instance, DST models, network models, and many representational models share the “dynamic” property that a small change in the input may have a large effect on the output (which is generally believed to be a property of the brain). Within the broad computer metaphor of the brain as an information-processing system, network models and representational models are primarily about specifics of the information process, whereas DST models are primarily about the system as such. DST models are about how systems develop over time and, in this sense, they would actually fit better in the narrow computer metaphor of the brain as a system that goes serially from one state to the next (like clocks and steam engines did in previous brain metaphors).

To push the latter point further, the formal DST concept of ‘attractors’ (i.e., states a system ends up in for many inputs) seems popular among DST adherents in cognitive science to suggest that an input triggers a competition between rivaling cognitive states, but this is a misleading suggestion. Such a competition is inherent to network models and many representational models, but it is not reflected in DST

models. Once an input has been given, a DST model applies differential equations to proceed stepwise along a specific path in state space, i.e., without taking into account states outside this path.

At this point, it seems expedient to distinguish between neural processing and cognitive processing. Neural processing refers primarily to the physical processes neurons undergo, and cognitive processing refers primarily to the functional processes stimuli undergo. DST models seem better suited to model neural processing, and network models and representational models seem better suited to model cognitive processing. Of course, eventually, both types of processing will have to be captured in one formal framework. I agree with Sundqvist that, regarding this goal, network models seem to have a good hand with, say, DST describable activation flows that might reflect successive brain states resulting in activation patterns that might reflect cognitive states. Traditional network models, however, seem to lack a theoretical connection between the physical and the psychical that representational models might provide. The network in a traditional network model has a fixed structure in which, beforehand, all relevant bits of information and interdependencies have to be represented for all possible inputs. Furthermore, activation spreading through the network links represents the interaction between bits of information that themselves are stored in the network nodes. This raises questions. Where do the bits of information and interdependencies in the model come from, and how are the bits of information supposed to be represented in the brain?

In my representational simplicity approach, I postulate that the fixed neural network of the brain performs a standard process that encodes structural stimulus invariants (i.e., spatial regularities). By reorganizing neural traces, this process creates various functional networks representing bits of information and interdependencies relevant for only the input at hand. In my computational model of this process (van der Helm, 2004), the role of nodes and links in these functional networks differs from that in traditional network models: The nodes represent spatial relationships between bits of stimulus information that themselves are represented by the links. This implements the idea that links are not interactors between bits of information but are information carriers that interact in the nodes. This way—in my view—activation patterns reflect better both physical Gestalts and phenomenal Gestalts. Notice that this picture also leaves room for Gestalts given by synchronous activation patterns in functional networks in different parts of the brain.

References

- Koenderink, J. J. (1999). Brain scanning and the single mind. *Perception*, 28, 1181–1184.
- van der Helm, P. A. (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, 126, 770–800.

van der Helm, P. A. (2004). Transparallel processing by hyperstrings. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 10862–10867.

PETER A. VAN DER HELM
Nijmegen Institute for Cognition and Information
Radboud University Nijmegen
Montessorilaan 3, 6525 HR Nijmegen, The Netherlands
Email: peterh@nici.ru.nl