Australasian Journal of Philosophy
Vol. 84, No. 1, pp. 129–145; March 2006

BOOK REVIEWS

Preston, John, and Mark Bishop, eds., *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford: Clarendon Press, 2002, pp. xvi + 410, US$99.00 (cloth), US$39.95 (paper).

This anthology's 20 new articles and bibliography attest to continued interest in Searle's (1980) Chinese Room Argument. Preston's excellent 'Introduction' ties the history and nature of cognitive science, computation, AI, and the CRA to relevant chapters.

Searle ('Twenty-One Years in the Chinese Room') says, 'purely . . . syntactical processes of the implemented computer program could not by themselves . . . *guarantee* . . . semantic content . . . essential to human cognition' [51]. 'Semantic content' appears to be mind-external entities 'attached' [53] to the program's symbols. But the program's implementation must accept these entities as input (suitably transduced), so the program in execution, accepting and processing this input, *would* provide the required content. The transduced input would then be *internal* representatives of the external content and would be related to the symbols of the formal, syntactic program in ways that play the same roles as the 'attachment' relationships between the *external* contents and the symbols [Rapaport 2000]. The 'semantic content' could then just be those mind-*internal* relationships, thus syntactic. Searle disagrees: The CRA 'rests on two . . .logical truths . . .[S]yntax is not semantics. . . .[S]imulation is not duplication' [52]. However, denying these isn't *logically* inconsistent: semantic correspondence between domains *can* arise from symbol manipulation, as just suggested, and simulation *can* be duplication, at least in the special cases of cognition and information: *Pace* Searle, simulated information *is* real information (consider a photocopy of a book), because an item's informational content lies in its abstract structure; structure-preserving simulations thereby contain the same information [Rapaport 1988; Rapaport 2005].

Block ('Searle's Arguments against Cognitive Science') complains that mechanically executing a natural-language-understanding program doesn't *feel* like language understanding: '[W]hen you seem to Chinese speakers to be . . . discours[ing] . . . in Chinese, all you are aware of doing is thinking about what noises the program tells you to make next, given the noises you hear and what you've written on your mental scratch pad' [72]. This nicely describes the situation of novice second-language speakers: consciously and laboriously computing what their interlocutor says and how to respond. Does this non-native speaker understand the language? The non-native speaker (or Searle-in-the-room) might say 'no', while the native speaker says 'yes'.

The native speaker's judgment should prevail [Turing 1950; Rapaport 2000], because, as Hauser ('"Nixin" Goes to China') notes, one can understand without *feeling* that one understands. The CRA is based on a 'dubious' principle that

Australasian Journal of Philosophy
ISSN 0004-8402 print/ISSN 1470-6828 online © 2006 Australasian Association of Philosophy
http://www.tandf.co.uk/journals DOI: 10.1080/00048400600571935

'first-person disavowals of understanding' are 'epistemically privileged' [127–8]—dubious because it does not follow from 'Searle's seeming to himself not to understand' [128] that Searle does not really understand.

For Winograd ('Understanding, Orientations, and Objectivity'), there is *no* 'answer to ... whether the computer (or the Chinese Room) "understands" language' [80]; language 'isn't prepared' for this: 'We have clear intuitions that... pencil sharpeners... don't understand, and that human[s]... do. But the computer is a mix-and-match' [82]. Winograd notes, correctly, that there are many different senses of 'understand'. But, fixing a sense of 'understand', we can ask of a computer or the CR whether it understands *in that sense*. Our judgement should be based on the same criteria in both cases [Turing 1950]. You or the computer understand a newspaper editorial in one sense if you can answer standard reading-comprehension questions, in another if you 'get' the political undertones. The computer's understanding should be no different.

Simon and Eisenstadt ('A Chinese Room that Understands') claim to provide such precise tests for NLU, viz., translation ability, NL question-answering, and 'similar tasks' [95]. But their claim that '*a* computer... has been programmed... to understand' NL [95; my emphasis] underwhelms: they present *three* programs, *each* of which implements a *partial* theory of NLU [101]. Nevertheless, this is progress.

Where Simon and Eisenstadt (and Aleksander, below) use real computer programs to *contradict* the CRA, Bringsjord and Noel ('Real Robots and the Missing Thought-Experiment in the Chinese Room Dialectic') claim that 'real robots... *strengthen*' it [145]. '[Z]ombanimals' [145] are real robots 'displaying our external behavior' without consciousness' [157f; cf. 146]. We are supposed to conclude that Searle-in-the-room equipped with 'the entire system' appears to see things with understanding, but 'clearly' [164] does not. 'Clearly' needs clarification: Perhaps the equipment needed to *appear* to see with understanding would be no different from *really* seeing with understanding.

Winograd's 'mix and match' status is explored in Adam's and Warwick's chapters. For Warwick ('Alien Encounters'), 'simple human biases' underlie the CRA, in part because it's possible to have a conscious machine whose consciousness does not arise from a computer program—witness 'insect-like' robots. But Warwick offers no evidence that these robots are *not* computational, nor that what they do is not comput*able* [see Harnad, below].

Adam ('Cyborgs in the Chinese Room'), echoing Warwick, advocates blurring the human-machine boundary [322] separating the 'profane'/'unclean'/'them' from the 'holy'/'clean'/'us' [Douglas 1966]. AI seen 'as outside the human body' is 'profane and unholy' [326]. Boundaries produce marginal members, viewed as dangerous or powerful. '[A]n alternative reading of machines in our society... include[s] them as marginal members... offer[ing] an explanation of why they are potentially threatening ...' [326]. Actor-network theory [Latour 1987] is like the Systems Reply, blurring this boundary by considering 'the process of creating scientific and technical knowledge in terms of a network of actors ...where power is located throughout the network rather than in the hands of individuals' [331]. On the other hand, 'cyborg feminism' [Haraway 1991] blurs the boundary by definition, since a 'cyborg' is a 'fabricated hybrid of machine and organism' [334], appearing to be a version of the Robot Reply.

For Proudfoot ('Wittgenstein's Anticipation of the Chinese Room'), Searle-in-the-room's Chinese utterances are not speech acts [167]: 'asking in Chinese "Are you in pain?" when the speaker does not know Chinese' is 'a paradigm example of talking "without thinking"' [167, citing Wittgenstein 1989]. Surely, a native Chinese speaker *could* sincerely ask Searle-in-the-room if he is in pain. Could Searle-in-the-room sincerely *answer*? If sincerely asked in English, *I* sincerely answer by knowing what 'pain' means, introspecting to see if I'm in pain, then answering. But who or what introspects in the CR? To whom or what should the interrogator's uses of 'you' refer (and Searle-in-the-room's uses of 'I')? If 'you' is the system (Searle plus Chinese instruction book), where would the pain (if any) be located, and how would it be sensed?

Rey ('Searle's Misunderstandings of Functionalism and Strong AI') may have a key to handling Proudfoot's problem: The instruction book must 'relate Chinese characters not only to one another, but also to the inputs and outputs of the *other* programs [that the 'Computational-Representation Theory of Thought' (CRTT)] posits to account for the *other* mental processes and propositional attitudes of a normal Chinese speaker' [208]. However, Rey also says, 'There's no reason whatever to suppose that the functional states of a *pain program memorizer* are the same as those of *someone actually in pain*' [214]. Rey's chapter, rich in ideas and bearing detailed study, claims that the CRA is irrelevant to CRTT [203], because the Turing Test is behaviouristic, concerned only with external input and output, and not 'committed to... a Conversation Manual Theory of Language' [207f], since a *lot* more is needed than a 'Berlitz Conversation manual' [208; cf. Rapaport 1988]. *Several* instruction books are needed, corresponding to interacting modules for various cognitive abilities, probably being executed in parallel, hence by more than one inhabitant of the Room (see Taylor, below). In this situation, possibly what's happening inside the room *would* be functionally equivalent to normal Chinese understanding. To the extent that the TT doesn't care about this, too bad for the TT and the CRA.

Harnad ('Minds, Machines, and Searle 2') defines 'computationalism' via three 'tenets': '(1) Mental states are just implementations of (the right) computer program(s) [which must] be *executed* in the form of a dynamical system' [297]. But rather than 'Computationalism [being] the theory that cognition is comput*ation*' [297, my italics], it should be the theory that cognition is comput*able* [Rapaport 1998]. While mental states are implementations of (the right) computer *processes*, *human* cognition could result from non-computational brain processes—as long as the behaviour is *also* characterizable computationally. Perhaps this is *part* of tenet (2): 'Computational states are implementation-independent' [297], implying that 'if all physical implementations of... [a] computational system are... equivalent, then when any one of them has (or lacks) a given computational property [including "being a mental state"], it follows that they all do' [298]. But equivalent in input-output behaviour, algorithm, data structures? Two such implementations might be *weakly* equivalent if they have (only) the same input-output behaviour; degrees of *strong* equivalence might depend on how alike the intervening computer programs were in terms of algorithms, subroutines, data structures, complexity, etc. Tenet (3) is that TT-indistinguishability is the strongest empirical test for the presence of mental states [298]. Harnad, echoing Rey, admits that this is input-output, i.e., weak, equivalence [299].

Taylor ('Do Virtual Actions Avoid the Chinese Room?') presents the CRA via slaves carrying out the Chinese NLU algorithm, suggesting an interesting variation on the Systems Reply: Here, no single person can claim, as Searle-in-the-room does, that he (or she) doesn't understand Chinese, yet Chinese is being understood. Thus, either *the entire system* (not any of its components) understands Chinese, or *nothing* does the understanding, despite understanding *happening*. Taylor meets Searle's challenge with 'a neurally based ... semantics' [270]. If Taylor means that one *neural representation* (of a word) is correlated with another *neural representation* (of an object), I approve. Unfortunately, he postulates that this is the site of Chomsky's [1965] deep structures, a theory no longer defended.

Bishop ('Dancing with Pixies') offers a weaker version of 'Putnam [1988]'s claim that, "every open system implements every Finite State Automaton (FSA)", and hence that psychological states of the brain cannot be functional states of a computer' [361]: 'over a finite time window, every open system implements the trace of a particular FSA. ... lead[ing] to panpsychism' and, by a *reductio*, 'a suitably programmed computer *qua* performing computation can [n]ever instantiate genuine phenomenal states' [361]. His argument is odd: For any Discrete State Machine (one capable of different output behaviour depending on its input), there will be several other machines, each with *fixed* input, such that each of these machines' output matches the DSM's output for the appropriate input. Then, for any *cognitive* DSM, 'we can generate a corresponding state transition sequence using any open physical system' [368]. But suppose that the cognitive DSM is 'collapsed' into several of these state transition sequences (presumably one per possible input). Choose one. Find an (arbitrary) 'open physical system' that has that same state transition sequence. It doesn't follow that that system is cognitive *just* because it does *part* of what the cognitive DSM does.

Haugeland ('Syntax, Semantics, Physics') explores the Systems Reply: Searle 'asks himself what it would be like if he were *part of* a mind that worked according to the principles that strong AI says all minds work on—in particular, what it would be like if he were the central processing unit' [379]. But 'neither the question nor the answer [viz., that the CPU does *not* understand Chinese] is very interesting' [379], since the CRA commits both part-whole and equivocation fallacies [382].

Coulter and Sharrock ('The Hinterland of the Chinese Room') assert: 'If computation requires intelligence, and ... can be done on machines, then, [Turing] ... thought, since machines can do computation they must possess intelligence' [184]. No: Turing [1936] argues that computation does *not* require intelligence; Turing 1950 argues for the *converse*.

Penrose ('Consciousness, Computation, and the Chinese Room') says that 'there must be *non-computational* physical actions underlying the brain processes that control our mathematical thought processes ... [and] that underlie our *awareness*' [236–7] because of his infamous Gödelian argument. But a human or a computer could use *two* formal systems, a proof-theoretic one and a corresponding model-theoretic one, both of which are syntactic (i.e., symbol-manipulation) systems, such that the former cannot prove some well-formed formula, while the latter determines that it is true.

Finally, others discuss variations on 'computation': Wheeler's 'Change in the Rules' concerns dynamical systems, Copeland's 'The Chinese Room from a Logical Point of View' discusses 'hypercomputation', and Aleksander's 'Neural Depictions of "World" and "Self"' considers 'neurocomputing'.

Despite Searle's sentiment that he's finished with the CRA, 'there is (still) little agreement about exactly how the argument goes wrong, or about what should be the exact response on behalf of computational cognitive science and Strong AI' [Preston: 47]. The CRA is an easy-to-understand and engaging argument around which a host of important philosophical issues can be approached. This book is a good place to explore them.

William J. Rapaport
University at Buffalo, SUNY

References

Chomsky, Noam 1965. *Aspects of the Theory of Syntax*, Cambridge MA: MIT Press.

Douglas, M. 1966. *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*, London: Ark.

Haraway, Donna, 1991. A Cyborg Manifesto: Science, Technology and Socialist-Feminism in the Late Twentieth Century, *Socialist Review* 80: 65 – 107.

Rapaport, William J. 1988. Syntactic Semantics: Foundations of Computational Natural-Language Understanding, in *Aspects of Artificial Intelligence*, ed. James H. Fetzer, Dordrecht: Kluwer Academic: 81 – 131.

Rapaport, William J. 1998. How Minds Can Be Computational Systems, *Journal of Experimental and Theoretical Artificial Intelligence* 10: 403 – 19.

Rapaport, William J. 2000. How to Pass a Turing Test: Syntactic Semantics, Natural-Language Understanding, and First-Person Cognition, *Journal of Logic, Language, and Information* 94: 467 – 90.

Rapaport, William J. 2005. Implementation Is Semantic Interpretation: Further Thoughts, *Journal of Experimental and Theoretical Artificial Intelligence* 17: 385 – 417.

Searle, John R. 1980. Minds, Brains, and Programs, *Behavioral and Brain Sciences* 3: 417 – 57.

Turing, Alan M. 1936. On Computable Numbers, with an Application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society*, Ser. 2; 42: 230 – 65.

Turing, Alan M. 1950. Computing Machinery and Intelligence, *Mind* 59: 433 – 60.

Wittgenstein, Ludwig, 1989. *Wittgenstein's Lectures on Philosophical Psychology 1946 – 1947*, ed. P. T. Geach, Chicago: University of Chicago Press.

Cullity, Garrett, *The Moral Demands of Affluence*, Oxford: Clarendon Press, 2004, pp. viii + 286, US$60 (cloth).

The last few years have seen a resurgence in the literature attempting to deal with the deep moral problems raised by the continued phenomenon of severe global poverty. This carefully argued and empirically informed book represents an impressive contribution to that literature. As its title suggests, the book focuses on the question of what beneficence demands of comparatively affluent people like us, given the existence of global poverty. Along with most moral philosophers, Cullity defends a 'moderate' position on this question: the requirements of beneficence are real and non-trivial, but not so extensive as to rule out the pursuit of ordinary projects and relationships. His arguments in support of, and against, this position are often novel and thought-provoking. Below, I shall flag some possible criticisms.

The book is divided into two main parts. In the first, comprising six chapters, Cullity presents and develops an argument for what he calls the 'Extreme Demand', essentially a version of the highly demanding argument based on a life-saving analogy made famous by Peter Singer. In particular, the argument's conclusion is that an individual is required to make contributions to life-saving agencies until the point at which her next contribution would itself constitute a large enough sacrifice to excuse her from contributing further. Cullity's development of this argument in Chapters 4 through 6 is sophisticated in its detail, and, given that he ultimately disagrees with it, highly sympathetic. Indeed, it was occasionally disconcerting to find myself expecting to see the appearance of the main counter-argument that I knew was going to appear in the second part of the book. However, the manner in which the book is structured allows Cullity to show just how potent the extremist argument is, despite its conclusion's lack of intuitive plausibility.

Of particular interest is his discussion, toward the end of Chapter 5, of the distinction between 'iterative' and 'aggregative' approaches to understanding the life-saving analogy that underpins the extreme demand. Basically, these terms correspond to different ways of determining the point at which the cost to a potential aid-giver constitutes a large enough sacrifice to excuse her from making a further contribution. According to the iterative approach, in making this determination one is not entitled to count previous contributions: the only relevant consideration is whether the cost in making this *particular* contribution is sufficiently onerous to excuse one from making it. On an aggregative approach, on the other hand, in making the determination an agent is entitled to count all contributions, past, present, and future, that fall within some reasonable time frame, e.g., one year. In effect, only an iterative approach to the life-saving analogy leads to the extreme demand. An aggregative approach leads to much less extreme demands insofar as it allows potential aid-givers to include past (and perhaps future) contributions in determining when they have given enough. As suggested, the main argument of the first part of the book can be read as supporting the iterative approach, but the second part consists of a sustained argument in favour of the aggregative approach.

I found the highlight of the first part of the book to be Chapter Three, 'Objections to Aid', notwithstanding the fact that it is slightly off the track of the main argument. The chapter contains an excellent summary of some of the best known objections to contributing to aid agencies, in particular the claim that there are good empirical reasons to think that doing so is actually counter-productive. Outside of academic philosophy, this is surely the best-known objection to providing aid, and Cullity is to be commended for giving it the careful attention that it deserves. After engaging with the objection through the words of some of its best known advocates, he then goes on to show that it nevertheless fails to refute the extreme demand. The chapter could easily serve as a reasonably self-contained and up-to-date introduction to the ethics of global poverty.

The second part of the book, 'Limits', comprises four chapters; it is followed by a final chapter that is an overview of the entire book. The key chapter in this second part is Chapter 8, 'The rejection of the extreme demand'. Cullity's approach here falls broadly into a class of arguments that proceed from the 'presuppositions' of beneficence, one that has had other well-known contemporary proponents including Bernard Williams. Employed against extremist accounts of beneficence, the basic idea behind 'presupposition' arguments is that extremism in some sense violates the very conditions or background suppositions of acting beneficently. For instance, in Williams's account, recognizing that I have reason to pursue my own interests and projects is a precondition of recognizing that I have reason to do anything. The implication is that extremist accounts of beneficence, insofar as they deny that there is any reason for agents to pursue their own interests and projects, are thereby unable to generate any reason for agents to promote the good of others.

Whether or not the inference at the heart of such an argument is sound, one may certainly doubt its key premise, namely, that extremist accounts are committed to denying that there is *any* reason for agents to promote their own interests and projects. For it would seem that such accounts are committed only to the weaker claim that, given the world we currently inhabit, the reason to promote one's own interests is usually trumped or outweighed by other reasons we have to help others in great need.

Cullity's version of the argument is somewhat different from the one just sketched. Given its overall importance to the book, it is worth summarizing in detail:

1. Extremist accounts require that we live 'altruistically focused lives', lives that severely constrict an agent's ability to pursue personal interests and projects except when these revolve around the goal of helping others.
2. It is the fulfilment of these very sorts of interests that makes life valuable, hence that provides the ground for many of the most compelling reasons we have to help others.
3. These legitimate interests may be contrasted with other interests in what it is *wrong* to have—interests that are *not* valuable and do not provide reasons for others to help those who have them.
4. Extremist accounts are committed to holding that personal interests and projects just *are* interests in what it is wrong to have, because such interests constitute a non-altruistically focused life, and it is wrong to live such a life.
5. (Conclusion) Extremist accounts are committed, absurdly, to denying that the fulfilments of a non-altruistically-focused life can ground morally compelling reasons to help others.

In line with what was suggested above, this argument can be questioned at premise 4, the claim that extremists are committed to holding that personal interests, projects, and relationships are *themselves* interests in what it is wrong to have. It can be argued, to the contrary, that the extremist needn't (and of course typically doesn't) claim that such interests are wrong in themselves, or that they have no legitimate weight, or that they can't ground reasons to help others. Rather, she may claim that such interests are *out*weighed by the urgent interests of others; and that, given the world we live in, they are outweighed to such an extent that we are required to live altruistically-focused lives.

Cullity is well aware of this sort of criticism, and he replies to it at a few different places. I can't do justice to his reply here, and I leave it to the reader to judge its plausibility. The gist of it, however, is that it is reasonable to say that the extremist is committed to holding that personal interests, etc. are themselves interests in what it is wrong to have. If this is right, then I believe he has produced a powerful argument against the extremist.

As noted, Cullity does not avoid the practical implications of his view, and Chapter 10 contains some general guidelines about how to deliberate about personal spending, as well as more specific proposals about what types of spending are likely to be (im)permissible. Some of these more specific proposals suggest a fairly demanding outlook (expensive purchases made purely for enjoyment are morally indefensible), while others seem quite permissive (spending resources on commitments of personal significance are acceptable). On the whole, I think it is fair to describe Cullity's practical suggestions as relatively demanding compared to most other moderates, though still well within the moderate range.

In conclusion, *The Moral Demands of Affluence* is an important book that is well worth reading. I recommend it highly to anyone with an interest in the ethics of global poverty. I also recommend that anyone who teaches applied ethics take a look at its third chapter.

Ramon Das
Victoria University of Wellington

Woodward, James, *Making Things Happen: A Theory of Causal Explanation*, New York: Oxford University Press, 2003, pp. viii + 410, US$65.00 (cloth).

The concept of causation plays a central role in many philosophical theories, and yet no account of causation has gained widespread acceptance among those who have investigated its foundations. Theories based on laws, counterfactuals, physical processes, and probabilistic dependence and independence relations (the list is by no means exhaustive) have all received detailed treatment in recent years—and, while no account has been entirely successful, it is generally agreed that the concept has been greatly clarified by the attempts. In this magnificent book, Woodward aims to give a unified account of causation and causal explanation in terms of the notion of a manipulation (or intervention, terms which can be read interchangeably). Not only does he produce in my view the most illuminating and comprehensive account of causation on offer, his theory also opens a great many avenues for future work in the area, and has ramifications for many other areas of philosophy. *Making Things Happen* ought to be of interest not only to philosophers of causation and philosophers of science, but to any philosopher whose concerns involve assumptions about the nature of causation, laws, or explanation.

The pre-theoretical notion of a manipulation is of a causal influence produced by an agent. Correspondingly, there are two traditional lines of objection against theories of causation formulated in manipulationist terms. Firstly, the concept appears anthropocentric (or at least agent-centric), threatening to introduce an unacceptable subjectivism (or at least agent-dependence) into what is supposed to be the paradigmatic objective relation. Secondly, since manipulation itself appears to be a causal concept, there is a worry of circularity. Woodward takes both lines of objection to count against earlier agency and manipulationist theories, and works hard to dissociate his theory from these. The difference is that Woodward does not offer a purported reduction of causation, but rather the explication of causal claims in terms of a notion of intervention that is itself defined as a particular kind of causal relationship. The circle is virtuous, since the theory shows how a great number of diverse causal concepts can be defined in terms of this particular causal concept. This strategy also avoids anthropocentrism, though by a kind of fiat—since the theory is non-reductive, a fortiori it does not reduce to anything anthropocentric. In taking this result to count against anthropocentrism Woodward simply rests on our pre-theoretical confidence in the objectivity of causation. In the final part of this review I will argue that his account itself gives reason to reassess this confidence; but to begin, I will give an overview of the theory to show some of its virtues, consequences and open questions.

The non-reductive approach to causation advocated by Woodward resembles and is influenced by the formal causal modelling frameworks pioneered by Judea Pearl and by the trio of Clark Glymour, Peter Spirtes, and Richard Scheines. (Woodward gives the best introductory treatment of this work I have seen). But where those theories take as a primitive a notion of causal mechanism, and define interventions in terms of these, Woodward takes the notion of an intervention as a primitive, and defines causal mechanisms (and other causal concepts) in these terms. This is a significant achievement, in two respects. Firstly, it sets the formal frameworks on stronger philosophical footing—for one thing, Woodward provides an account of the meaning of causal claims as embodied in these frameworks, the lack of which has been the focus of recurring criticism; for another, starting with interventions rather than mechanisms fits far more easily with the epistemological and methodological dimensions of causal explanation. Secondly, this inversion has the advantage that the path to a potential reductionist account of causation, in terms of agency, is made clear. Indeed, one of the great benefits of the book is that it brings together two traditions that have hitherto proceeded largely independently of one another—on the one hand a tradition originating in econometrics and experimental design, and continued in contemporary work on causal inference in computer science, which takes causal claims to encode claims about the results of hypothetical experiments; and on the other hand a philosophical tradition that attempts to analyse causation in terms of agency.

Starting with interventions leaves the question of their relationship to causal generalizations and laws. In Woodward's account it is the notion of invariance under interventions that plays the role laws of nature do in other theories—to distinguish between causal and merely accidental generalizations. Invariance under interventions holds when a particular generalization correctly captures the counterfactual relationships between two variables under a particular range of interventions. As Woodward notes, 'whether or not a generalization is invariant is surprisingly independent of whether it satisfies many of the traditional criteria for lawfulness, such as exceptionlessness, breadth of scope, and degree of theoretical integration' [17]. Independent, that is, because *weaker*—it may capture a generalization holding in quite particular circumstances, for quite particular interventions. This might appear *too* weak—a generalization only capturing the relationship between one or two possible interventions for some particular situation hardly merits the title—but the benefit is that we have a continuum, from minimal sorts of invariance all the way through to the ideally exceptionless invariance (invariance under all possible interventions) of the laws of physics. Indeed, Woodward considers laws 'as just one kind of invariant generalisation' [17]. Obviously, some explanations are more informative than others, and Woodward [18] proposes that explanatory depth, in the same way, can be analysed in terms of the degree of invariance that the explanations support. This is a lovely and intuitive way to characterize the difference between laws of nature and the laws of the special sciences, through to the sorts of everyday causal knowledge embedded in folk psychology. And it has the desirable consequence that we can see scientific knowledge as an elaboration and refinement of everyday causal thinking rather than taking the latter to involve implicit knowledge of the former [20], a point on which Woodward's account is clearly superior to rival models of explanation.

Nevertheless, there is a range of open questions concerning how the idealized definitions of causal concepts provided by Woodward can be mapped onto real world practices, the resolution of which is especially important given Woodward's insistence early in the book that a theory of causation needs to involve an epistemological aspect that makes causal knowledge accessible to ordinary agents. At first glance, definitions such as his **DC** (Direct Cause [55]) do not satisfy this desideratum—in order to make a true direct-causal claim, for example, we would need to have the ability to perform (or otherwise counterfactually ascertain the result of) an intervention on a system with all other variables also held fixed by intervention. Needless to say, this is not something we ordinarily do, or are even in all cases capable of doing. Similar questions arise for other definitions Woodward provides, leaving a rich area of investigation for cognitive scientists and like-minded philosophers to explore. The benefit of the formal apparatus employed is that it makes these questions precise. Indeed, the precision Woodward's approach gives to questions of conceptual clarification is remarkable. A quite minimal apparatus is required to both elegantly describe and then diagnose our judgments concerning classic cases such as cancellation of total causal impact along multiple causal routes, failures of causal transitivity (dog bite [57 – 9], falling boulder [79 – 81]), and various purported counterexamples to counterfactual analyses of causation (chestnut smasher, [67ff], thirsty traveller, [77ff], trumping [81 – 2]). Many controversial cases in the literature have continued to be debated even when all sides agree about the relevant counterfactual dependencies, and the apparatus Woodward uses both explains the various intuitions involved, and how they can be reconciled within a manipulationist framework.

Moreover, the framework involves no metaphysical claims whatsoever, simply employing a distinction between individuals, types, and variables used to represent those individuals and types. The causal relata on Woodward's view [111 – 14] are simply any particulars that can be manipulated—whether these be facts, events, tropes, or any other metaphysical candidate you wish to plug in (manipulation implies that these particulars be capable of taking different values; thus Woodward suggests that *variables* are the best way to characterize the causal relata). The lack of metaphysical claims masks, however, the degree to which the framework might help metaphysical debates in other areas. Take mental causation. Central to contemporary debates in this area has been the exclusion problem, where the possibility of alternative explanations for behaviour in terms of physics and in terms of belief-desire psychology are supposed to generate metaphysical worries about the efficacy of mental states qua mental. According to Woodward's account of causation, these explanations simply don't compete—each is framed in terms of a different variable set and is a bona fide causal explanation just in case the relevant counterfactuals are true. There is no *causal* sense in which physical explanations exclude or otherwise diminish mental explanations, though there might be further interesting questions concerning the relations between the two. In fact, Woodward [147] says that his account does not imply that all causal explanations are backed by exceptionless laws; and that if this is in fact the case, it will be an interesting *empirical* discovery that has no bearing on the truth of higher-level causal claims. So if we accept Woodward's account of causation, causation will not be a metaphysical concept driving reductionist arguments in philosophy of mind—there being no sense in which physical explanations are more causal than any other form of explanation. Reductionist arguments will have to find some other way of privileging the physical.

This line of argument is available because on Woodward's account, causal claims are relative to the specification of variables, both in the sense of which variables are included in the set [55 – 6] and of how fine or coarse grained the specifications are [378 – 9 n. 20]. This obviously leaves open a range of further questions, analogous to familiar issues in philosophy of science: How do we select a variable specification (model)? Can this be done independently of causal claims? Are some models better than others, and if so, on what grounds? Can models be compared independently of causal claims? Every model will make claims that can be objectively tested by performing the relevant hypothetical experiments they embody (or in cases where manipulation is practically impossible, by otherwise evaluating the relevant counterfactuals)—the questions here, however, arise in comparing models each of which is empirically adequate. While these are all pressing questions, it is nice to have them disentangled from questions about causation per se, which on Woodward's account can play no role in their resolution.

In closing, I will make some brief comments on Woodward's claim that his theory avoids the agent-dependency of earlier manipulationist theories. Woodward [85–91] makes a distinction between what he takes to be the agent-independent patterns of counterfactual dependence that constitute the 'objective core' of causal claims, and the agent-dependent pragmatic features of those claims which he takes to consist in the selection of those counterfactuals that represent 'serious possibilities'. Nevertheless, we do make causal claims in the absence of the practical or physical possibility of performing the associated manipulations, a point which has been another common line of criticism of agency theories of causation. The solution Woodward adopts to this problem [131–2] is to weaken the strength of possibility required for interventions, so that it is only required that they be logically or conceptually possible, and that we have some means of evaluating the relevant counterfactuals (Woodward is critical [118ff] of projectivist views of causation, so it is interesting to note that this itself amounts to a form of projectivism). The cleanest form of logically possible intervention is simply to have the state of the world at the time of the change miraculously become such that (only) the change has occurred. (Such a change trivially meets Woodward's requirements for an intervention; formally, we can suppose the required intervention variable be *God's choosing to make it so*.) And the simplest means of evaluating the relevant counterfactuals is to use laws of nature—after all, these are the fundamental invariances. But these constraints are too weak for the purposes of recovering causal claims, since they will license counterfactuals in both the past and future directions in time. (This sort of claim is often put in terms of the time-symmetrical nature of the laws of physics, but here requires simply that the laws serve equally well for purposes of retrodiction and prediction).

There are several different strategies that might be used to recover the temporal asymmetry of causation here, but I think that it suggests that agent-dependency is not so easily evaded. We can put the question to Woodward in the same form as he puts a very similar question to his rival accounts. Woodward asks, of those who propose that causation is a disunified cluster concept, why we shouldn't abandon our concept of causation in favour of some revised version, causation* [93], and with respect to Lewis's similarity metric, why we shouldn't exchange it for another metric and a corresponding notion of smausation [137]. So the question for Woodward is—why this pattern of counterfactual dependence and not another? Or put slightly differently, why one sort of counterfactual antecedent and not another? Why is it that one sort of counterfactual is the sort that we can use for the purposes of manipulation, and not other sorts? Again, given that any variety of counterfactual meeting the criteria of an intervention will give us a variety of manipulation, why is it only some subset of these that we are interested in? Why shouldn't we abandon counterfactual for counterfactual*, especially if counterfactual* will enable us to cause* past events? The answer, I think, is that we can't, in fact, bring about counterfactual* antecedents (at least in all cases we know of)—but this is in part a fact about the sorts of agents we are.

Early in the book [28], Woodward suggests that the demand for a reductionist account of causation 'virtually forces one' to an anthropocentric conception of causation. And the train of thought underlying much of the resistance to such a reduction seems to be that anthropocentrism is equivalent to subjectivism, and the insistence that whatever causation is, it can't be subjective. The mistake here is in supposing agent-dependence to be equivalent to subjectivism—the fact that we can't bring about counterfactual* antecedents might be agent-dependent, but it is certainly not subjective. Here as elsewhere, *Making Things Happen* helps to focus the issues in a way that allows theoretical progress; it deserves to form an axis around which future debates in causation and explanation revolve.

Brad Weslake
University of Sydney

Dennett, Daniel, *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*, Cambridge MA: MIT Press, 2005, pp. xiii + 199, US$28 (cloth).

In *Sweet Dreams* Dennett presents himself as a Lockean under-labourer, 'removing some of the rubbish that lies in the way to knowledge'. The rubbish in question is identified in the book's subtitle: 'philosophical obstacles to a science of consciousness'. Dennett's central claim is that the intuitions and thought experiments that dominate philosophical discussions of consciousness are hampering the scientific study of consciousness. Removing these obstacles will reveal that consciousness poses no hard problems and raises no explanatory gaps; indeed, it will reveal that a mechanistic explanation of consciousness is not just possible but is 'fast becoming actual' [7].

Although written with Dennett's characteristic panache, *Sweet Dreams* is less than kind on its reader. Dennett's Jean Nicod lectures (chapters two through five) are sandwiched between various papers on consciousness, but rather than being presented as stand alone papers they are arranged as chapters. There is too much repetition between them for this arrangement to be successful. Numerous claims and indeed entire paragraphs reappear verbatim. Even in these environmentally conscious times I wouldn't have thought that rubbish-removal entailed quite so much recycling.

Also likely to try the patience of some readers is Dennett's characterization of his opponents. Zombiephiles rely on an intuition that is 'almost entirely arational, insensitive to argument or the lack thereof' [22], while scientists who suspect that consciousness might not succumb to current scientific methods have been 'tempted' or 'blackmailed' into holding these views [134]. In light of these comments one might have expected *Sweet Dreams* to be packed with arguments. Not so. There are arguments here and there, but it is not always easy to find them among the metaphors (consciousness as fame; consciousness as cerebral celebrity; consciousness as fantasy echo) and stories (the Tuned Deck, Mr. Clapgras, and the very entertaining Indian Rope Trick). In fact, Dennett is rather pessimistic about the ability of reason to resolve the qualia wars:

> the tempting idea that there is a Hard Problem is simply a mistake. I cannot prove this, and some who love the Hard Problem find my claim so incredible that they admit, with some hilarity, that they can't take it seriously. So I won't make the tactical error of trying to dislodge with rational argument a conviction that is beyond reason.
>
> [72]

It is not hard to have some sympathy with this pessimism: it *is* difficult finding argumentative traction on the issues that divide qualiaphiles from qualiaphobes. Is this, as Dennett's implies, because the qualiaphiles are allied with the forces of irrationality? I think not. Rather, I suspect that it has more to do with the fact that there is insufficient common ground between Dennett and his foes for argument to get much of a toehold.

Dennett's themes in *Sweet Dreams* are familiar enough: zombies are conceptually incoherent; Mary doesn't learn anything when she sees red for the first time; there are no intrinsic phenomenal properties; heterophenomenology is unavoidable. Although little of this material constitutes an advance over positions that Dennett has previously articulated, there is much here to give the qualiaphile pause for thought. But rather than tread these (well-trodden) paths, I want to examine the claim with which Dennett frames his discussion: zombiephilia poses obstacles to the science of consciousness.

What exactly are the obstacles that Dennett has in mind? In what way(s) could the metaphysics of consciousness be an impediment to the science of consciousness? Dennett says little that explicitly addresses these questions, but he does offer the reader a very interesting *hint* as to what he takes the problem to be. He points out that zombiephiles—'reactionaries', as he refers to them—have occasionally suggested that we should look to fundamental physics in order to explain consciousness [8 – 9]. Now, as Dennett himself knows, embracing the Hard Problem has not led consciousness scientists to forsake cognitive neuroscience in favour of fundamental physics. But perhaps his point is that it *should* have: if zombiephiles really understood their own position that would realize that it is inconsistent with the attempt to understood consciousness using the tools of the cognitive neuroscientist.

The standard qualiaphile response to this argument has been articulated by David Chalmers in a number of places, and runs roughly as follows. Although phenomenal properties are fundamental physical properties, they give rise to—or take part in—relations with non-fundamental physical properties. These derivative relations are accessible via the methods of cognitive neuroscience, and with lots of work and a bit of luck we will be able to work our way down from the derivative relations that hold between phenomenal states and higher-level physical states to the primitive relations that hold between phenomenal states and fundamental physical states.

Will this work? Well, maybe; but given the explanatory autonomy of the special sciences this does seem like an awfully ambitious research programme. I would have thought that interesting cognitive-level generalizations involving consciousness are unlikely if phenomenal properties are fundamental physical properties. Of course, the issues here are very tricky and sorting them out will have to wait for another occasion. What does seem clear at this stage of the debate is that the qualiaphile has a more challenging job making sense of the current methodology of consciousness studies than do those who locate phenomenal properties at a biological level.

Dennett is not just an under-labourer, he is an *optimistic* under-labourer: philosophical obstacles notwithstanding, a mechanistic explanation of consciousness is 'fast becoming actual' [7]. Dennett's money is on the global workspace approach to consciousness [131]. What it is for content to be conscious is for it to occur within a global workspace, wherein it can be broadcast to the system as a whole [132].

But hold on—isn't the Cartesian Theatre Dennett's *bete noir?* How is it possible to reconcile Dennett's multiple drafts conception of consciousness with his enthusiasm for the global workspace? What is a global workspace if not that dreaded arena wherein 'it all comes together'?

Well, perhaps there is *some* daylight between the global workspace and the Cartesian Theatre. Instead of thinking of a global workspace as a 'consciousness module', one might think of it as a colourful way of referring to those processes, modular or not, that allow content to be globally available for the rational control of thought and action. Understood in this way, Dennett's endorsement of the global workspace account amounts to nothing more than a restatement of his claim that Block's notion of access-consciousness is the only concept of consciousness that there is.

But this solution has a sting in its tail: if it is right, then the global workspace account is not, contrary to advertising, an *empirical* account of consciousness. Rather, it is trivially entailed by Dennett's functional analysis of the concept of consciousness—an analysis that is essential to his rejection of the possibility of zombies. No wonder that Dennett is optimistic about the prospects of the global workspace account—it seems to fall right out of his analysis of the concept of consciousness!

Dennett's account of the concept of consciousness also problematizes his discussion of the function of consciousness. Dennett presents recent research on unconscious perception as giving us an account of the function of consciousness. He suggests that the upshot of this data is that consciousness enables one to use information strategically [141]. But again, is this news? Not if one is operating with a purely functional analysis of the concept of consciousness, as Dennett does. We do not need sophisticated psychophysics to tell us that consciousness is necessary for the strategic control of information if this result is entailed by our very analysis of the concept. And unless (something very much like) it is so entailed, then we face the zombie threat.

One of the questions facing Consciousness Studies at the moment concerns the role that philosophers should play in proceedings. Although I suspect that this is not Dennett's considered view, the over-riding impression he leaves one with is that the best the philosopher can do is get out of the consciousness kitchen and leave the cognitive neuroscientists to it. It would be a pity if this attitude were to take root.[1]

Tim Bayne
Macquarie University

Deutscher, Max, *Genre and Void: Looking Back at Sartre and Beauvoir*, Aldershot: Ashgate, 2003, pp. xxxii + 268, US$79.95 (cloth).

Max Deutscher's study, *Genre and Void*, sets out to reinvigorate some of Jean-Paul Sartre's and Simone de Beauvoir's ideas 'so as to keep them in motion as part of contemporary thinking, not rendered *passé* by structuralism and post-structuralism' [ix]. Such revitalizing is not meant as a return to the heyday of existentialism, a philosophical era that 'no writer has the power to recreate' [x]. Instead, Deutscher encourages a discussion not of, but *through*, the existentialist themes raised by these thinkers, especially as relating to the alienating effects of certain forms of individualism facilitated through often gender-coded and conflictual promises of personal fulfilment. The individual is drawn into inherently antagonistic visions of social existence that oscillate between calls for loving intimacy, and public achievement

[1] Thanks to Uriah Kriegel, Jordi Fernandez, and especially David Chalmers for helpful comments on an earlier version of this review.

through participation in our society's dreams of salvation via 'technologically mediated humanity' [218].

An immediate advantage of Deutscher's approach is that his examination is not limited to the traditional routes of individual psychology and ideology critique. Instead, he pursues a discourse-analytical trajectory, broadly conceived, to capture the different strata of communal life in which the meanings of private and public life are produced, circulated, and authorized. Since discourses multiply and work at cross purposes in any community, Deutscher problematizes certain postmodern accounts of *gender* construction and identity politics, exemplified by Judith Butler's work, as he intimates at the very end of his study [254 n. 16]. Rather than refuting her position on performativity, however, he takes the double-voiced solution of his title, using *genre* instead of *gender*, to stress how gender-coding works among that for other social types.

Generally, Deutscher does not claim his mobilization of Sartre's and Beauvoir's ideas for a contemporary audience as either an extention of, or alternative to structuralist and post-structuralist thought. Yet his primary passages on the *systemic* features of certain philosophical discourses [5, 24, 53, 141] appear indebted to Julia Kristeva whom he mentions only in passing [216 n. 24]. Similarly, Deutscher's extensive use of the *field* metaphor in describing the 'power of thought' [48; cf. 46–8, 54, 56, 244] would seem to call for an acknowledgment of Pierre Bourdieu's models for discursive authority and possible resistance. His references to discursive 'territory' [58, 174, 193] and 'philosophical personages' [251], highly reminiscent of Gilles Deleuze and Félix Guattari, receive only an early gloss [xxx].

Instead, *Genre and Void* draws strongly on Michèle Le Dœuff's critique of Sartre. Toward the end of Deutscher's investigation he concludes: 'We have seen how Beauvoir is subverting his [Sartre's] phenomenology of an *inevitable* antagonism between "I" and "Other", both by an ethical appeal ("generosity") and an importation of issues of economy, social structure and political factors into a philosophical understanding of what goes on in constructing an "Other"' [250].

Deutscher's 'Introduction' announces this course: '[i]n Part I, Beauvoir's work gets more close attention than Sartre's does' [xiv], while Sartre's will take centre stage especially in Parts II and III, followed by a co-engagement with the contemporary critiques of Beauvoir by Luce Irigaray and Le Dœuff, reflected on Sartre. Yet from the outset, even Sartre's metaphysics are viewed through the critical lens of Le Dœuff.

In Chapter Two, Beauvoir has her say [15–28]. Under the rubric of 'Dreams, Fears, Idols', Deutscher astutely profiles Beauvoir's account of myths (e.g., '*woman as devourer*' [27]) as systems that 'thrive on contradiction' [24]. Yet here, where Beauvoir is arguably at her most poignant, Deutscher shifts gears into Sartre's rather different account of 'bad faith'. The reader will have to wait until Chapter Eight [173ff.] for Beauvoir's text to be engaged in any substantial detail, with one important exception.

In Chapter Three, 'Bound to be Free', we find a significant hint at Beauvoir's working notion of *solidarity*, presented in sharp contrast to the limits imposed by Sartre's 'demoralis[ing]' [58] outlook: While Beauvoir's analysis is keyed to 'that very work of solidarity that achieves rather than decrees a universal free consciousness', according to Deutscher, 'the need for solidarity with the oppressed must be grafted on to Sartre's existential ontology' [ibid.]. 'Universal free consciousness' may already overstate Beauvoir's case, but this reviewer would have welcomed a more extensive discussion of what it means to 'graft' a conception of solidarity on to Sartre's own ontology.

In this regard, Deutscher's Derridian 'deconstruction' [xxxi; cf. 215], as the purported catalyst of a 'critical attitude' [46] overtaxes the scope of language in a way that leaves his argument vulnerable to Hans Albert's long-standing critique of a 'pure hermeneutics' that extends the text-metaphor to all of reality. In other words, Deutscher's primary case for the power of 'play' [47; cf. 158–9] and 'parody' [59] rides mostly on a pantextualist reduction, which construes all the cultural-political forces at work in a societal setting as only so many aspects of language.

Somewhat contrary to his introductory remarks about Derrida's influence on his present project, Deutscher would object that '[w]e cannot make Sartre contemporary by "doing a Derrida" upon him, or by making a Derrida of him' [43]. Upon closer inspection, however, this is largely what happens in *Genre and Void.*

The deconstructionist's recommendation to satirize or 'make light of' [xii] symptoms of oppression to subvert them is vacuous, unless there are criteria for identifying those symptoms from the first. Instead, Deutscher seems to use phrases such as 'urgent injustice' [60] or '[t]he love affair of Church and Fascism' [214] as unproblematic or self-evident. The main suggestion about Sartre's conception of 'absolute freedom'—the (theological?) trope of 'sacrific[ing] part of [one's] conceptual investment' [60]—then, remains as suggestive as it is unexplained.

Deutscher seems generally reluctant to support Beauvoir's 'subversion of Sartre's phenomenology', which leads his discussion at times to approach self-contradiction, especially in his treatment of sexual consciousness. He readily concedes that '[f]or the most part the differences between Beauvoir and Sartre about sexual connection as an ideal of intimacy could not be greater' [187]. Now Deutscher's earlier exposition of Beauvoir's work on contradiction-driven myths is restaged as a blueprint to help theory negotiate 'the possibilities and meanings of heterosexuality and homosexuality as explicit themes' [ibid.], for which Sartre's use of 'he' as a universal subject offers no promising equivalent at all. He does not, however, concede a solution to Beauvoir: '[t]he ill effect that remains in Beauvoir's writing, all the same, is that sexual feelings and conduct emerge in the text only at an impersonal level. These are the abstractions that she shared with, borrowed from, or lent to Sartre' [ibid.]. *Genre and Void* thus finds Beauvoir mired in abstractionism, even as her work is compatible with contemporary models of discursive systems and territorialized meanings, including Beauvoir's purported sensitivity to 'material realities' [181] or Deutscher's subsequent comments on her 'social materialism' [252].

Ultimately, then, *Genre and Void* combines a series of what appears to be conclusive theoretical objections to Sartre with a rhetoric of redemption. The study insists that, against all odds, Sartre's existentialist project is still worthwhile considering for contemporary debates over genre, gender, and myth, especially for what seems to be a *concrete* sexualized phenomenology.

Deutscher's text opens the door to new appropriations of Sartre and Beauvoir. Drawing from materialist semiotics, fantasy theory, and certain strands of theological realism, such phenomenology could tap, for example, new understandings of 'naturalized divinity' through a continued critical exchange between Beauvoir and Irigaray. Thus understood, Beauvoir could emerge as a pathbreaking pioneer of French *Existenz*-philosophy, rather than French 'existentialism'. Alternately, *Genre and Void* could point a way toward redeeming Sartrean thought.

Deutscher has not used his own heuristic optimally: a discourse-analytical framework could work more historically at the territorializing effects of concrete genres at particular sites. Beauvoir's project could thus take shape, e.g., as part of an open intra- and postwar constellation, where historical materialism, different Marxisms, and linguistic mysticism were reshuffled into alternative conceptions of community and community-grounding rituals. Closest to Beauvoir, those rituals were explored, for instance, by the so-called College of Sociology in Paris in the late 1930s, as Denis Hollier has documented.

Accordingly, *Genre and Void* retains its value as a refutation of Sartre, while it insists on the possibility of his rehabilitation. As such, it underrates Beauvoir's contribution to a theologically sensitive phenomenology and its productive power for creating a new genre, or genres, of sexual solidarity. Yet Deutscher points the way toward a re-evaluation of Beauvoir's struggle with philosophical 'fatalism' [182] and the technologies of myth, affirming the value of her work for the existential heritage within present-day semiotic theory.

Markus Weidler
University of Auckland