Minds and Machines (2020) 30:513–532
https://doi.org/10.1007/s11023-020-09533-8

GENERAL ARTICLE

Deceptive Appearances: the Turing Test, Response-Dependence, and Intelligence as an Emotional Concept

Michael Wheeler[1]

Received: 7 May 2020 / Accepted: 29 July 2020 / Published online: 12 August 2020

Abstract

The Turing Test is routinely understood as a behaviourist test for machine intelligence. Diane Proudfoot (Rethinking Turing's Test, *Journal of Philosophy*, 2013) has argued for an alternative interpretation. According to Proudfoot, Turing's claim that intelligence is what he calls 'an emotional concept' indicates that he conceived of intelligence in response-dependence terms. As she puts it: 'Turing's criterion for "thinking" is…: x is intelligent (or thinks) if in the actual world, in an unrestricted computer-imitates-human game, x appears intelligent to an average interrogator'. The role of the famous test is thus to provide the conditions in which to examine the average interrogator's responses. I shall argue that Proudfoot's analysis falls short. The philosophical literature contains two main models of response-dependence, what I shall call the *transparency* model and the *reference-fixing* model. Proudfoot resists the thought that Turing might have endorsed one of these models to the exclusion of the other. But the details of her own analysis indicate that she is, *in fact*, committed to the claim that Turing's account of intelligence is grounded in a transparency model, rather than a reference-fixing one. By contrast, I shall argue that while Turing did indeed conceive of intelligence in response-dependence terms, his account is grounded in a reference-fixing model, rather than a transparency one. This is fortunate (for Turing), because, as an account of intelligence, the transparency model is arguably problematic in a way that the reference-fixing model isn't.

Keywords Artificial intelligence (AI) · Imitation game · Machine intelligence · Response-dependence · The Turing Test

✉ Michael Wheeler
m.w.wheeler@stir.ac.uk

[1] Division of Law and Philosophy, University of Stirling, Stirling FK9 4LA, UK

1 The Shape of Things

In a 1948 report to the *National Physical Laboratory* entitled 'Intelligent Machinery', Alan Turing claims that intelligence is what he calls an 'emotional concept'. What does he mean by this? And what does that tell us about the place and function, in Turing's own thinking, of his famous but controversial test for machine intelligence, the so-called Turing Test? In what follows, I shall offer answers to both of these questions. My strategy for doing so will be to develop a critical treatment of an important, state-of-the-art paper in Turing scholarship by Proudfoot (2013). Departing from the established mainstream understandings of Turing's view (more on which later), and by focusing centrally on Turing's aforementioned characterization of intelligence as an emotional concept, Proudfoot argues that he pursued a *response-dependence* approach to thought and intelligence. (Turing didn't systematically distinguish between testing for thinking and testing for intelligence. Proudfoot follows him in that, and so shall I.) In wielding the notion of response-dependence, Proudfoot is drawing on a recent philosophical literature in which the animating intuition is that there are certain concepts—the response-dependent ones—whose extensions (the set of things that count as falling under the concept) are, in some important sense, determined by particular sets of human responses (for canonical statements of the position, see e.g. Johnston 1989; Pettit 1991; Wright 1988). As a first-pass statement of this idea, consider the concept of red. If red is a response-dependent concept (which many philosophers think that it is), then, if an entity appears red to a particular kind of subject in certain specified conditions, then that entity is red. In other words, appearing red to a particular kind of subject in certain specified conditions is sufficient for being red. Similarly, according to Proudfoot, on Turing's response-dependence account of intelligence, it is an entity's capacity to appear intelligent, to a particular kind of subject, in certain specified conditions, that suffices for that entity to be intelligent. On this understanding of Turing, the role of the Turing Test is to provide the appropriate conditions for revealing the all-important responses of the relevant subject.

In this paper, I shall argue that Proudfoot's analysis of Turing's view falls short. Although I shall not question—indeed I shall present new evidence in support of—the claim that Turing pursues a response-dependence approach to intelligence, I shall articulate and defend a version of that idea that differs significantly from the one proposed by Proudfoot. Following a brief statement of where the target analysis by Proudfoot fits into the debate over how to understand the Turing test, plus an introduction to the response-dependence approach itself, the argument of this paper unfolds as follows. The modern philosophical literature contains two main models of response-dependence, what I shall call the *transparency* model and the *reference-fixing* model. At a key moment in her treatment, Proudfoot resists the thought that Turing might have endorsed (implicitly of course) one of these models to the exclusion of the other. In effect, her claim is that Turing's own species of response-dependence did not distinguish between them. However, I shall argue that once the distinction between these models is

brought into proper view, Proudfoot's official reluctance here is revealed to be problematic. In the first instance, this is because once one examines the details of her analysis, it transpires that she is, *in fact*, committed to the claim that Turing's account of intelligence is grounded in one of these models rather than the other. More specifically, she is committed to the claim that Turing's account of intelligence is grounded in a transparency model of response-dependence, rather than a reference-fixing one.

This would be no more than a relatively minor expository clarification, were it not for the fact that the evidence also indicates that it is the diametrically opposite interpretation of Turing which is correct. In other words, the evidence suggests that Turing's account of intelligence is grounded in a reference-fixing model of response-dependence, rather than a transparency one. As I shall go on to explain, this is a fortunate result (for Turing), because thinking of intelligence on the transparency model arguably confronts a serious philosophical challenge, one that simply doesn't arise if we think of intelligence on the reference-fixing model. And what this means is that, on my interpretation of Turing, his account of mind and intelligence is in far better shape than it is on Proudfoot's interpretation. Thus this paper aims to deliver not only a critical response to Proudfoot's analysis, but also a distinctive and improved positive interpretation of Turing on machine intelligence, one that boosts the likelihood that Turing was right. As a second advantage of the reference-fixing approach, I shall suggest that it increases the resilience of the Turing Test in the face of a certain common style of objection.

It is worth emphasizing at the outset that the dispute at the heart of this paper is more than a turf war amongst Turing-obsessives over an inconsequential historical wrinkle in his view. In addition to his ground-breaking technical innovations within the field of computing in general, Turing made an important contribution to foundational philosophical issues in what we now call artificial intelligence (AI), most obviously in his 1950 *Mind* paper, 'Computing Machinery and Intelligence', the paper that is routinely taken to be the *locus classicus* of the Turing Test. This contribution concerns the question of what it would be for a machine to think or be intelligent, so it pivots centrally on the notions of thought and intelligence themselves. As a consequence, to understand and evaluate Turing's place in the philosophy of AI, we need to get straight about what he meant by such terms. Moreover, it is worth registering the fact that, in the contemporary philosophical debate over the scope of response-dependence, it remains an open question whether it is correct to think of psychological phenomena such as thought and intelligence in response-dependence terms. Here it is surely tempting to see a prominent theory of mind such as Dennett's intentional systems theory as encompassing a version of the claim that psychological terms are response-dependent concepts. After all, if, as (Dennett 1981, p. 72, original emphasis) famously claims, '*all there is* to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy', where the intentional strategy is, at root, the practice of attributing psychological states to things, then the extension of any psychological concept (the set of entities which are, for example, genuine believers) is, in an important sense, determined by particular sets of human responses (such as finding some observed behaviour predictable as a result of attributing the concept in question). Dennett's view continues

to be influential in cognitive science and AI, which is just one reason for holding that the prospects for a response-dependence approach to thought and intelligence is a topic of contemporary scientific, as well as philosophical, concern. And if, as I shall argue, Turing was a response-dependence theorist of a particular stripe about thought and intelligence, and if getting straight now about the precise shape of the view he advocated plausibly enables the Turing Test to take on a newly resilient status within AI, that surely strengthens the hand of the response-dependence approach by enhancing its usefulness as a framework for conceptualizing and identifying machine intelligence. Thus although the present treatment is openly historical in its primary orientation, meaning that a full exploration of its wider contemporary implications is outside its scope, those implications surely extend beyond 'mere' historical book-keeping.

2 Looking for Machine Intelligence

In 'Computing Machinery and Intelligence', Turing proposes his famous test as a replacement for the question 'Can machines think?' (Turing 1950, p. 441), a question that he himself judges to be 'too meaningless to deserve discussion' (*ibid*. 449). As Turing sees things, to answer the question 'Can machines think?' directly, we would plausibly need to know what the words 'machine' and 'think' mean, but any attempt to provide such definitions would, at least if the aim is to reflect their normal usage, propel us into an 'absurd' statistical polling of the population in order to harvest their views on the matter, a project that would inevitably run aground on problems of imprecision and ambiguity (*ibid.* 441). As things turn out, what, for Turing, actually replaces the question 'Can machines think?' is an alternative question that, as we shall soon see, should be asked in relation to the Turing Test. But, first things first. What exactly is the test? Although 'Computing Machinery and Intelligence' is routinely treated as the canonical presentation of the idea, at least three versions of the test appear within Turing's treatments of, as we would now say, AI (Turing 1948, 1950; and in Braithwaite et al. 1952). Here is the version that will concern us in this paper. It is sometimes referred to as the unrestricted imitation game and it combines elements from more than one of Turing's own descriptions.

In stage one of the test, a human interrogator sits in front of a computer terminal, typing in questions or, more generally, conversational prompts for responses (which is what makes the game unrestricted) and reading responses from the screen, responses which come from two remote sources. In the text on the screen, these remote sources are labelled simply as A and B. Now, A and B themselves are located out of direct sensory contact with the interrogator, in a different room. As it happens, A is a man and B is a woman. The challenge confronting the interrogator is to correctly identify the source with the gender, that is, to identify A as the man and B as the woman. This task is rendered non-trivial by the fact that while it is the goal of the woman (B) to help the interrogator make the right identification, it is the goal of the man (A) to fool the interrogator into making the wrong identification.

How does this man-imitates-woman game bear on the question of machine intelligence? That's where the second stage of the test comes in. The game is now

modified as follows. The role of A (the deceiver) is to be taken by a carefully programmed and adequately powerful (in terms of storage capacity and speed) digital computer. This is the kind of machine that Turing thought was in the right ballpark to exhibit intelligence, in virtue of the fact that digital computers were explicitly designed to carry out operations standardly performed by human computers (Turing 1950, p. 444). The role of B (the truth-teller) is to be taken by a human being, with the gender of that participant presumably no longer intended to be salient (although this is a debated issue; for discussion, see Genova 1994; Traiger 2000; Copeland 2000; Piccinini 2000; Moor 2001). Thus, when Turing returns to it later in 'Computing Machinery and Intelligence', stage two of the imitation game looks like this: 'Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by [a human being]?' (Turing 1950, p. 448). This is the alternative question that replaces Turing's initial question, 'Can machines think?'.

So, the computer's goal in the imitation game is to deceive the interrogator into misidentifying it as a human being, with the overall outcome of the test judged by reference to whether the percentage of correct identifications on the part of the interrogator, over some agreed-to-be-adequate length of time, is worse than, the same as, or better than, the percentage of correct identifications on the part of the interrogator in the man-imitates-woman game. If the computer in stage two of the test is as good, or better, at deceiving the interrogator into misidentifying it as the human being, as the man in stage one is at deceiving the interrogator into misidentifying him as the woman, then the computer has successfully passed the Turing Test and has met what Turing calls his 'criterion for 'thinking'' (Turing 1950, p. 443).

In the corridors of philosophy and AI, the Turing Test is routinely depicted as the principal driver for the conclusion that Turing was some sort of behaviourist about intelligence. More specifically, it is often claimed, citing the Turing test as evidence, that Turing held that it is logically sufficient for an entity to be intelligent, or to be a thinker, that that entity exhibits a certain behavioural profile, whatever might be going on in that entity's innards (see e.g. Searle 1980). At first sight, this is not an unreasonable interpretation. As just one piece of evidence, in a 1952 BBC radio debate involving Turing and three others, Turing's sometimes colleague, the mathematician, codebreaker and computer pioneer Max Newman, says: 'If I have understood Turing's test properly, you are not allowed to go behind the scenes and criticize the method, but must abide by the scoring on correct answers, found reasonably quickly' (in Braithwaite et al. 1952, p. 496). Within that discussion anyway, Turing doesn't object to Newman's interpretation. Still, behaviourism is not the only route. What might be described as the other mainstream alternative in the literature portrays the Turing Test as specifying the conditions under which a machine's behaviour may justifiably count as the grounds for an induction regarding intelligence (Moor 1976, 2001). On this view, if the machine passes the test, we have good inductive grounds to conclude that it is intelligent. If this is the right way to go, then Turing is not a behaviourist (since the test-passing behaviour isn't logically sufficient for the presence of intelligence), but it seems fair to say that, whatever the

merits of the inductive view, it hasn't gripped the imagination of philosophers or AI researchers, in the way that the behaviourist interpretation has.

This is where Proudfoot (2013) offers us a third, potentially game-changing option, by arguing that Turing advocated a response-dependence account of intelligence, according to which it is an entity's capacity to appear intelligent, to a particular kind of subject, in certain specified conditions, that suffices for that entity to be intelligent. Call this *the response-dependence interpretation of Turing on intelligence*, henceforth *RDI*. To understand and evaluate RDI, we will of course need to say more about who the relevant subject in question is and about the conditions in which that subject's responses matter. It is in meeting these requirements that we will see how, within the response-dependence framework, the contribution of the Turing Test itself is reconceived.

3 The Response-Dependence Interpretation of Turing on Intelligence

Let's begin by focusing on the pivotal notion of response-dependence itself. As mentioned already, the basic intuition behind response-dependence is that there are certain concepts whose extensions are, in some important sense, determined by particular sets of human responses. To put some flesh on this, here is a general schema for response-dependence. (In fact, this schema, as I am about to formulate it, puts rather too much flesh on the intuition, even though it reflects a common way of elaborating it, but we'll come back to that shortly.)

> Take a property F, an entity x, a specified kind of observer O, and a specified set of conditions C. If our concept of F is a response-dependent concept, then the following will hold: it is a priori true that x is F if and only if O judges x to be F in C.

Here are two immediate clarifications that could themselves be prompts for discussion, but which I shall simply note. First, the term 'judging' should be heard in an essentially non-committal way, so as to allow for a range of different sorts of responses, given different cases of response-dependence. Such responses might include, among other things, having certain sensations, having certain feelings or emotions, and making considered evaluations. Secondly, it is incumbent upon the fan of response-dependence to specify O and C such that they do not produce what Wright (1989) calls 'whatever-it-takes' formulations. These are formulations that succeed in making the key claim a priori true, but only in a trivial way, say by specifying O as those observers whose relevant responses accurately track F things, and C as whatever conditions are favourable for accurately tracking F things.

If Turing held that intelligence is a response-dependent concept, then, on the basis of our suggested schema, he would have adhered to the following principle: it is a priori true that x is intelligent if and only if O judges x to be intelligent in C. But now, with our sights set on Proudfoot's analysis, we need to make some adjustments. First, as far as RDI is concerned, Proudfoot does not, in her text, explicitly mention the commitment to the a priori truth of the key embedded claim. Now, as it happens, the a priori status of that claim is an issue to which we shall return much later in this

paper. For now, however, in the interests of moving swiftly to a statement of Proudfoot's view, I shall simply suppress that element of the schema. Secondly, whereas our general schema is formulated in terms of a set of necessary and sufficient conditions, Proudfoot formulates RDI only as a set of conditions that, when met, are sufficient for the presence of thought and intelligence. As she puts the fundamental idea: 'On a response-dependence approach, it is an agent's capacity to appear intelligent (to a normal subject in specified conditions) that suffices for intelligence' (Proudfoot 2013, p. 404). Although this elimination of the expected necessity requirement as part of a response-dependence approach is not explicitly justified *as such* during Proudfoot's analysis, it is, I think, bound up with two other parts of her interpretation of Turing: her endorsement of the point that Turing was not attempting to provide a *definition* of intelligence, but rather a criterion for the presence of the phenomenon (*ibid.* 394); and the observation that it is reasonable to doubt whether Turing held that passing his test was necessary for the presence of thought (*ibid.* 398). In any case, all we need to do is follow suit.

With the noted adjustments in place, here's what we have so far as a statement of what, according to RDI, Turing believes: x is intelligent if O judges x to be intelligent in C. The next steps are to specify the relevant O and the relevant C, and to say what 'judging' amounts to, in the particular case we care about. In effect, that's what Proudfoot does, in the following, fuller statement of RDI: 'According to the response-dependence interpretation… Turing's criterion for "thinking" is…: x is intelligent (or thinks) if in the actual world, in an unrestricted computer-imitates-human game, x appears intelligent to an average interrogator' (*ibid.* 402). This formulation is immediately augmented by the following footnote: '[a]nd if the interrogator in that game is taken in no less frequently than the interrogator in a man-imitates-woman game' (*ibid.* 402, note 54). This more detailed formulation of RDI tells us what we need to know.

The specified conditions in which subject responses should be tested (C) are twofold: (i) the relevant judgments are those that are made in an unrestricted, imitation game form of the Turing Test, as described earlier this paper, and (ii) that game must take place in the actual world. The first condition plays an important role in ensuring a fair playing field for the machine in the test. For example, versions of the test that diverge from its dual-player format, by involving sequentially presented single interviewees, some of which are machines and some of which are human beings, result in a bias effect in which human contestants are regularly misidentified as machines, presumably because the judges over-compensate due to a strong desire not be hoodwinked by a computer (see Copeland 2004a, pp. 488–489). The second condition prevents any slide into discussions about logically possible, but not possible in the actual world, machines that, in our fertile imaginations, pass the Turing test, but which we might be nervous about regarding as intelligent. Enormous (beyond-reality-enormous, that is) look up tables, of one sort or another, are popular examples (Proudfoot 2013, p. 400).

The relevant observer, O, is identified as an 'average interrogator' in the imitation game. The word 'average' here needs to be given a specific interpretation, which, in modern parlance, we might express as being that the interrogator should not be an expert in AI. This respects Turing's insistence that the interrogator should 'not be

an expert about machines' (in Braithwaite et al. 1952, p. 495). The reasons for this restriction on the choice of interrogator will be important later, so further discussion of the issue will be deferred until then.

Turning to what is designated by the term 'judges' in this context, the simple answer is revealed if one lines up the phrase '*O* judges *x* to be *F*' from our generic response-dependence schema with the phrase 'x appears intelligent to an average interrogator' in Proudfoot's detailed formulation of RDI. This immediately gives us the claim that the average interrogator judges the target entity to be intelligent when x appears intelligent to her. However, this is where the aforementioned footnote becomes significant. This cites the now-familiar success condition for the computer in the Turing Test, that is, that the interrogator in the computer-imitates-human game should make statistically as many or more relevant misidentifications as were made by the same interrogator in a prior man-imitates-woman game. As far as I can see, this should be understood not as adding anything substantial to RDI as formulated in the main text (which explains why the information is placed in a footnote), but rather as an articulation of what it means here for an entity to appear intelligent to the average interrogator (what it means for the average interrogator to judge the target entity to be intelligent). Meeting the stated success condition is just what 'appears intelligent' means, in the context of the game.

Put all this together and RDI is plausibly complete. Whereas a response-dependence account of the concept of red might naturally appeal to the red-sensation responses of a normal subject operating in good lighting conditions, RDI will appeal to the judgments of the average interrogator in the unrestricted imitation game, played in the actual world. And now we can see why RDI, as Proudfoot develops and defends it, encompasses a specific and, in comparison with the canonical, behaviourist interpretation of Turing described earlier, a notably different role for the Turing Test. As Proudfoot (2013, p. 395) observes, on the behaviourist interpretation of Turing, it remains something of a mystery just why he constructed the test as the imitation game at all. Why make the responses of an interrogator central to the test when one could simply check to see if the machine is capable of executing a series of behavioural tasks? On RDI, however, no such mystery is established, since the interrogator's responses necessarily perform a perspicuous and important function in relation to determining the extension of the concept of intelligence. More specifically, according to RDI, the multi-faceted contribution of the Turing Test, in the form of the unrestricted imitation game, is to provide the following: a specification of the conditions under which the relevant subject's responses should be examined; an articulation of what the relevant responses are; and part of the profile of the relevant subject, with the remainder provided by Turing's extra requirement that the interrogator be 'average'.

4 Two Models of Response-Dependence

So far, so good. But now it's time to make the transition to a more critical register. To do this, we need to revisit, yet again, our understanding of response-dependence. There are, in fact, two main models of response-dependence in the philosophical

literature. These are sometimes distinguished by asking whether the focus of the view is on the concept under consideration or on the property of which it is the concept. For example, when one gives a response-dependence account of red, one might ask whether that is intended as an understanding of our concept of red or as an understanding of the property of being red. Proudfoot herself adopts this way of drawing the distinction at issue, and, in a brief remark, immediately dismisses any consideration of it in relation to RDI as inappropriate. She states, without further argument, that it is 'anachronistic' to 'ask whether Turing proposes a response-dependence understanding of the property, rather than the concept, of intelligence—*in fact his remarks suggest both*' (Proudfoot 2013, p. 308, my emphasis). Therefore, although Proudfoot clearly endorses the thought that there is a perfectly respectable distinction to be drawn between two models of response-dependence, she officially rejects the further thought that Turing embraced one of these models to the exclusion of other.

In order to assess this (as we shall see) surprisingly important moment in Proudfoot's analysis, we need to begin by getting clearer about the very distinction that Proudfoot has in mind, that is, the distinction between the two main models of response-dependence. In truth, it is misleading to separate these models using the simple 'concept versus property' locution, since, as Crispin Wright (personal communication) has pointed out to me, and in harmony with the way I introduced the notion of response-dependence earlier in this paper, *both* models think of certain human responses as, in some important way, *determining the extension of the concept in question*. In other words, *both* models propose that certain human responses have significant implications *of some sort* for which things in the world fall under the concept in question and so instantiate the property of which it is the concept. What really separates them is not whether the focus is on the concept or on the property, but rather their different ways of conceiving the relationship between the relevant set of responses and the property in the extension-determining process.

Consider first the model of response-dependence that was provisionally articulated as focusing on the property. Put more precisely, the key feature of this model is that the target concept is transparent with respect to the property of which it is the concept. Let's call this model of response-dependence *the transparency model*. On the transparency model, there is nothing more to instantiating the property in question than generating certain specified responses in certain specified subjects under certain specified conditions. Accordingly, if red is a response-dependent concept, then there is nothing more to being red than generating red sensations in normally sighted subjects under normal lighting conditions. And because, on this view, it is sufficient for something to instantiate a target property that it be such as to generate a certain specified response in the relevant subject in certain specified conditions, having the specified response necessarily acquaints the responding subject with the presence of the property in question. Of course, what I have just given is a very general exposition of the model on offer, and the completion of the view would involve significant philosophical work and may demand some tinkering with the details (see e.g. the different versions of this approach pursued by Johnston and Wright, as mentioned earlier). However, for our purposes here, the general picture will do.

Now consider the other main model of response-dependence—the one that was previously articulated as focusing on the concept. The key feature of this model is that the target concept may be opaque with respect to the property of which it is the concept. Our responses play an essentially provisional, but nevertheless still important, role in determining the extension of the concept. However, that role is an indicative or reference-fixing one, where what is indicated or referred to is some presumed, non-relational, response-independent essence—what we might think of as the underlying nature of the property in question. Ultimately, it is this underlying nature that determines the extension of the concept. Let's call this second model of response dependence *the reference-fixing model*. To illustrate the model, consider a natural kind such as water. We start out with a concept of water delineated by a set of human judgments or responses, in this case a combination of sensory appearances such as feels wet, is colourless and tasteless, and so on. This set of responses provisionally determines an extension by, in a sense, attaching it to a presumed, underlying physical essence (as it turns out, H_2O), which we understand to be the final determiner of the extension of the concept. To capture the kind of response-dependence in play here, one might say that, on this model, water is *introduced* as whatever it is that accounts for the characteristic set of responses that normal subjects experience in its presence. Notice immediately that, in contrast with the transparency model, there is now more to instantiating the target property than generating a certain specified response in certain specified subjects under certain specified conditions. So there is more to being water than generating the highlighted set of sensations in normally experiencing subjects under everyday sensory access conditions. Water has an underlying essence with which we are not directly acquainted, but which is indicated by the responses in question. Once again, what I have just given is a very general exposition of the model on offer, and the completion of the view would involve significant philosophical work and may demand some tinkering with the details (see e.g. the work of Pettit mentioned earlier). However, for our purposes here, the general picture will do.

With our two models of response-dependence in hand, it is now possible for us to state RDI in two different forms, according to whether the response-dependency in question is of the transparency or the reference-fixing variety.

> *RDI-transparency*: Turing holds that there is nothing more to a computer being intelligent than being judged as such by an average interrogator in an unrestricted computer-imitates-human game, where 'being judged as intelligent' just means 'achieves at least the same number of misidentifications on the part of that interrogator as the male participant achieves in a prior man-imitates-woman game'.

> *RDI-reference-fixing*: Turing holds that where the average interrogator in an unrestricted computer-imitates-human game judges a computer to be intelligent (makes at least the same number of computer-is-human-misidentifications in that game as man-is-woman-misidentifications in a prior man-imitates-woman game), that judgment provisionally determines the extension of the concept of intelligence to include the computer. It does so by presuming that

the computer instantiates an underlying essence of intelligence that ultimately determines the extension of the concept of intelligence.

Obviously, more needs to be said about what is meant by 'underlying essence of intelligence' in RDI-reference-fixing. We'll come back to that issue later.

Returning now to Proudfoot's analysis of Turing, how should we respond to her claim that Turing proposes a response-dependence understanding of both the property and the concept of intelligence? To answer that question, we need to restate her claim in light of our refined way of making the distinction in question. At first sight, the obvious restatement might seem to be that Turing proposes that we should understand the concept of intelligence in terms of both the transparency model and the reference-fixing model simultaneously. After all, structurally speaking, Proudfoot's thought is that there are two models of response-dependence and, with respect to intelligence, Turing advocates both. But now something has gone wrong: for although the two models could apply simultaneously to different concepts—perhaps the transparency model works for red and the reference-fixing model works for water—the very same concept cannot, at the same time, be both transparent and opaque with respect to the property of which it is the concept. So, that cannot be the charitable way to restate Proudfoot's claim. The correct resolution of this situation, I suggest, is for Proudfoot to be understood as claiming that Turing adopts the transparency model. For while both models begin with an elucidation of a target concept in terms of certain responses in specified conditions, it is on the transparency model that there exists the closer or more direct relationship between the accompanying target property (the property of which the target concept is a concept) and the aforementioned responses and conditions. On RDI-transparency, there is nothing more to a machine being intelligent than its being judged as such by the identified subject in the specified conditions. As one might crudely put it, according to RDI-transparency, when one gives an account of the concept of intelligence in terms of certain responses in specified conditions, one thereby says all there is to say about the property of intelligence. That isn't the case according to RDI-reference-fixing. On RDI-reference-fixing, the instantiation of the property of intelligence is ultimately determined by which entities in the world realize some presumed response-independent underlying essence of intelligence. So it is the transparency model that offers us the natural scaffold for the thought that Turing adopted a response-dependence account of both the concept and the property of intelligence.

The interpretative conclusion just drawn already suggests something important, namely that Proudfoot's version of RDI depicts Turing as pursuing one model of response-dependence rather than the other. This means that her official dismissal of the idea of applying the distinction in question to Turing's thinking about intelligence in order to ask which of the two models he embraced is, by her own lights, misplaced. However, one would be forgiven for complaining that the rather swift moves just made reek of a technical knock-out and, as Fodor (1987) once quipped, nobody wants to win anything by one of those. So we need to work harder. It is time to investigate for ourselves some of the very passages from Turing's writings on machine intelligence on which Proudfoot bases her case for RDI. Over the next two sections, I shall argue that a careful reading of those passages indicates not only

that Proudfoot is, *in fact*, committed to the claim that Turing's account of intelligence is grounded in a transparency model of response-dependence rather than a reference-fixing one, but that this is a mistaken interpretation of Turing, because Turing's account of intelligence is actually grounded in a reference-fixing model of response-dependence rather than a transparency one.

5 Intelligence as an Emotional Concept

In 'Intelligent Machinery', Turing claims that the concept of intelligence is what he calls an *emotional concept* or *emotional idea*. This claim occurs during Turing's response to a brace of reasons that he identifies as explaining a widespread resistance to the very idea of machine intelligence. Those reasons are that humankind will refuse to admit the possibility that it 'will have any rivals in intellectual power' and the 'religious belief that any attempt to construct [intelligent] machines is a sort of Promethean irreverence' (Turing 1948, p. 410). Turing's first reaction is to state that objections based on these reasons 'do not really need to be refuted', because they are 'purely emotional' (*ibid.* 411). In other words, whatever 'being purely emotional' means, if an objection falls under that description, then that is justification enough to ignore that objection. This will be important later. Having lodged this response, however, Turing proceeds to make an additional point that seems to pull in precisely the opposite direction, namely that the objections under consideration 'cannot be wholly ignored, because the idea of 'intelligence' is itself emotional rather than mathematical' (*ibid.* 411). Turing's thought, as I understand it, is that because the target idea—that of machine intelligence—features a concept—that of intelligence—that is itself an emotional concept, it is inevitable that objections to the target idea based on similarly emotional concepts, objections that might otherwise be ignored on that very basis, have some traction. Clearly we need to understand what it is for a concept to be emotional, and to do that we need to understand the distinction that Turing draws between emotional ideas and mathematical ones.

At the end of 'Intelligent Machinery', Turing states that intelligence is an emotional concept because the 'extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration' (*ibid.* 431). Here Turing distinguishes between two sorts of factors: (i) the properties of the object under consideration, and (ii) other factors that determine our judgments about the object in question, namely our own state of mind and training. Building on this while extrapolating from the specific case of intelligence, and putting just a few words into Turing's mouth, here is what I take to be Turing's distinction between emotional ideas and mathematical ones. A concept is mathematical if and only if, in judgments regarding whether that concept is instantiated in an object under consideration, only factors of sort (i) are relevant. A concept is emotional if and only if, in judgments regarding whether that concept is instantiated in an object under consideration, factors of sort (ii) weigh as heavily as those of sort (i).

As Turing explains things, factors of sort (ii)—the sort of factors whose equal role in judgments regarding the extension of the concept in question determines that

that concept is emotional in nature—include the judging subject's state of mind and her training. So now what do these terms signify? Given our everyday use of the phrase 'state of mind', it might seem that Turing is straightforwardly directing our attention to an affective component in the relevant judgments. That fits, of course, with Turing's designation of the concepts in question as 'emotional'. But while the reasons that Turing identifies as being behind the resistance to machine intelligence—namely human chauvinism and religion—credibly have affective dimensions, it is worth stressing that they are more like deep background perspectives on the world that only allow the world to open up or make sense in certain ways. In any case, Turing moves beyond the affective altogether in his explication of factors of sort (ii) when he includes the judge's training as such a factor. 'Training' refers to how much the relevant judging subject knows about the subject matter or domain under examination. Turing's thought here is striking, and I shall return to it below, but his general idea is that the more a judging subject is able to explain and predict the behaviour of an entity, the less likely it is that that subject will attribute intelligence to that entity.

One way to appreciate, for ourselves, how the foregoing material might function as a driver for some form of RDI is simply to read the term 'regard' in the phrase 'extent to which we regard something as behaving in an intelligent manner' as *rightly* regard, which, in context, doesn't seem unreasonable. It then becomes entirely plausible that, on Turing's view, whether or not an entity is intelligent (i.e., whether it falls under the extension of the concept of intelligence), depends, in an important way, on the state of mind and training of the subject making the judgment (i.e., on a subject's responses in certain conditions). So it is tempting to see a clear route from thinking of intelligence as an emotional concept to thinking of it as a response-dependent concept. Of course it is not until the fully worked out version of RDI is on the table that we can see how this might really work, and it is credit to Proudfoot's ingenuity that she manages to show how the Turing test allows us to fill in the all-important details for the machine intelligence case. As we have seen, these details include: who counts as the relevant judging subject (the average interrogator, where average means not an expert about machines); what the appropriate conditions are in which her judgments should be made (the unrestricted computer-imitates-human game played in the real world); and what counts as judging (the comparison between the percentage of misidentifications in the man-imitates-woman and the computer-imitates-human rounds). But even though there is a sense in which Proudfoot is manifestly going beyond what Turing actually says here, it is grist to her mill that an early formulation of the Turing test appears, in 'Intelligent Machinery', immediately after the unpacking of what it means for a concept to be emotional (*ibid.* 431).

So, let's agree that RDI, in some form, undoubtedly has legs. Are they transparency legs or reference-fixing legs? That is the question. In the next section I shall argue that, in tension with her official position, Proudfoot's analysis ends up portraying Turing as an advocate of a transparency model of response-dependence in relation to the concept of intelligence. I shall then argue that his actual model is a reference-fixing one.

6 Turing and Response-Dependence: Which Model?

Proudfoot claims that, for Turing, no computational account of the inner processes that generate behaviour could ever establish the presence of intelligence. As she puts it: 'Evidence of a machine's inner computation is evidence of (to use Turing's words) "the properties of the object under consideration"—in modern terminology, the machine's response-independent properties. However, it is not evidence of any response-dependent property, such as intelligence' (Proudfoot 2013, p. 403). This indicates that Proudfoot's Turing is working with the transparency model of response-dependence. Why? Because although Proudfoot's claim makes perfect sense on the transparency model, it is at odds with the reference-fixing model. On the transparency model, since there is nothing more to a computer being intelligent than being judged as such by an average interrogator in an unrestricted computer-imitates human game, the inner computational profile of the machine simply does not bear on the issue. On the reference-fixing model, however, an inner computational profile could indeed provide evidence of intelligence, by being elected as the underlying nature of the property to which the relevant responses of the interrogator in the imitation game are attached in fixing the reference of the concept of intelligence, that is, by being the underlying nature that ultimately determines the extension of that concept. Thus, on the reference-fixing model, just as H_2O provides evidence of the presence of water, so a particular computational profile (whatever the right one may be—classical, connectionist, predictive processing or otherwise) may provide evidence of the presence of intelligence.

It is worth pausing to emphasize the point that, when it comes to intelligence, the response-independent underlying nature that we are taking Turing to advocate as the ultimate extension-determiner is not a low-level physical structure, as it is in the case of water and H_2O, but a computational profile, a higher-order functional kind that, in line with a well-established understanding of the metaphysics of computational systems, might be multiply realized in a range of different physical systems. No doubt this is a good thing, for at least three reasons. First, the history that we teach undergraduates in introductory philosophy of mind classes emphasizes the transition from the type–type mind-brain identity theory to functionalism precisely as a way of respecting the point that, while we might still hope to find a shared cognition-delineating functional profile across thinkers, we gave up the hope of ever finding a shared cognition-delineating physical profile. Just remember those Martian thinkers who are made of different physical stuff to us. And, of course, conceiving intelligence in terms of computational states and processes is just a way of implementing functionalism. Secondly, advocating a functional/computational underlying nature for intelligence is one way of respecting the widely held view in AI—a view that, as we have seen, fuels scepticism regarding the credentials of the behaviouristic version of the Turing test—that, whatever the behaviour, not any old inner profile will do as a wellspring of genuine intelligence. And thirdly, introducing the possibility of a computationally specified underlying essence of intelligence blocks a perhaps tempting line

of thought which goes like this: because it is highly unlikely that intelligence has an underlying physical nature, the only response-dependence option available is the transparency model. If the relevant underlying nature is a functionally/computationally characterized one, the failure to find a physical one is beside the point. Anyway, what matters right now is this: the fact that the underlying nature in view when we focus on intelligence is to be functionally, rather than physically, characterized doesn't affect the basic shape of the view on offer. And the upshot of that is that it's only on the transparency model that response-independent properties provide no evidence of intelligence. Therefore, Proudfoot must be attributing that model to Turing.

It must be admitted that the foregoing interpretative conclusion flies in the face of some of what Proudfoot officially says. For example, she appeals, in part, to Pettit's (1991) understanding of response-dependence to save Turing from an unpalatable in-the-eye-of-the-beholder anti-realism about intelligence, even though Pettit is the chief architect of the reference-fixing model (Proudfoot 2013, p. 405). And, as part of the same resistance, she notes explicitly the part of Turing's definition of intelligence as an emotional concept (see earlier) which guarantees that, when it comes to determining the extension of that concept, properties of the object under consideration—properties which, as we have seen, include the machine's computational profile—matter just as much as the responses of the relevant subject as determined by her state of mind and training (*ibid.* 405). And that, as we have just seen, is in tune with the reference-fixing model in a way that it isn't in tune with the transparency model. The diagnosis of why such appeals persist in Proudfoot's analysis is that they are a symptom of her aforementioned explicit refusal to choose between the two models in relation to Turing's views. Given the argument just developed, the onus is, I think, firmly on Proudfoot to justify why she has access to these elements of the reference-fixing model.

Having concluded that Proudfoot attributes the transparency model to Turing, I shall now suggest that she has made a mistake in so doing. In fact, we have just registered one consideration that tells in this direction, namely that the reference-fixing interpretation of Turing makes better sense of the way in which 'the properties of the object under consideration' figure in his definition of emotional concepts. But there is more to be said, beginning with a switch in attention back to the alternative way of expressing the dual dependence of emotional concepts, that is, by concentrating again on the fact that, for Turing, the extensions of such concepts are determined as much by the judging subject's state of mind and training as they are by the properties of the object under consideration.

First consider training. As we know, Turing claims that the more successfully a judging subject, faced with the behaviour of an entity, is able to explain and predict the behaviour of that entity—the better trained the subject is in the domain of performance—the less likely it is that she will attribute intelligence to it. Thus, in the 1952 radio debate, he says:

> As soon as one can see the cause and effect working themselves out in the brain, one regards it as not being thinking, but a sort of unimaginative donkey work. From this point of view one might be tempted to define thinking as

> consisting of 'those mental processes that we don't understand'. If this is right, then to make a thinking machine is to make one which does interesting things without our really understanding quite how it is done. (In Braithwaite et al. 1952, p. 500)

Observations such as these lead Turing to conclude that, in the imitation game, the interrogator should 'not be an expert about machines' (*ibid.* 495). This in turn leads Proudfoot, in response-dependence mode, to specify that the relevant judging subject in RDI is the 'average interrogator'. Now, here is something obvious but crucial. When Turing says that the human chauvinist or someone with a certain religious commitment will, because of their state of mind, refuse to countenance the possibility of machine intelligence, he clearly believes that person to be in error. And surely he thinks the same about the trained AI expert who, because she knows plenty (too much) about programs and algorithms and so can, at least in a general, way, explain and predict the behaviour of a machine, declines to attribute genuine intelligence to it. The question, then, is how to explain such errors, on a response-dependence model.

On the transparency model, the judging subject's 'error' is presumably to be handled entirely by the fact that the AI expert is not the average interrogator. The expert's responses are ruled to be outside the set that, according to the relevant response-dependence schema, determine the extension of the concept. Put another, deliberately tendentious way, the domain-expert is wrong simply because she disagrees with the domain-novice. There is surely something intuitively uncomfortable about this way of proceeding. The AI expert is wrong, but it seems false to say that that's because she's unqualified to judge or simply because she disagrees with the domain-novice. That is presumably not Turing's intention. Fortunately, for RDI, the reference-fixing model supplies us with an alternative strategy. On that model, our capacity to criticize the domain-expert is explained by the fact that, even though the responses of the average interrogator are used provisionally to fix the reference of the concept of intelligence, something else is elected to play the role of ultimately determining the extension of that concept, something about which even the AI-expert may be wrong, namely the response-independent computational profile that constitutes the underlying nature of intelligence. So, by interpreting Turing's views on intelligence according to the reference-fixing model, we can have our response-dependence cake and yet eat it with a healthy side plate of plausibility when it comes to making good sense of his treatment of how training affects the credentials of certain subject-responses in the Turing test.

7 Two Advantages of the Reference-Fixing Model

In this section, I shall suggest that, when it comes to the phenomenon of intelligence, the reference-fixing model of response-dependence enjoys two advantages over the transparency-based alternative. This means that my Turing, who applies the reference-fixing model, is in better shape than Proudfoot's Turing, who (ultimately) applies the transparency model. The first of these advantages concerns a potentially

dangerous circularity that confronts the transparency approach but not the reference-fixing approach. To be clear, my aim is not to establish, beyond all reasonable doubt, that the circularity in question is decisively damaging, only that it's a threat. So I shall not explore potential responses to it from the perspective of the transparency model. It is enough for my purposes that having to deal with that threat is a headache that, on my interpretation of Turing, he doesn't have.

On the transparency model, there is nothing more to instantiating a target property than generating certain specified responses in certain specified subjects under certain specified conditions. Thus, canonically, there is nothing more to being red than generating red sensations in normally sighted subjects under normal lighting conditions. Notice that here there is not even a whiff of circularity. The account of what is sufficient to instantiate the property of being red is given in terms of certain specified perceptual experiences (red sensations), not in terms of being red. Now consider the example of intelligence. Applying the transparency model, one is encouraged to give an account of what is sufficient to instantiate the property of being intelligent in terms of judgments (here, those made by the appropriate interrogator in the Turing test) that are themselves instantiations of intelligence. Here there does seem to be a circularity in evidence and, given the risk that the account on offer will be rendered uninformative, it is not obviously benign.

The reference-fixing model avoids this alleged circularity, because its generic structure involves a gap, one that doesn't exist on the transparency model, between the specified responses and the property of which the target concept is a concept. This gap is opened up by the fact that the responses that provisionally fix the referent of the concept are only contingently connected to the response-independent underlying nature that ultimately determines the extension of the concept. In effect, that's just another way of developing the point that the reference-fixing model characterizes the responses in question as provisionally determining the extension of a concept by attaching it to a presumed, underlying essence which we understand to be the final determiner of that extension, such that what is ultimately sufficient for an entity to instantiate the property in question is that it realizes that presumed underlying nature.

Notice that there is no longer any cast-iron guarantee that the specified reference-fixing responses, whatever they may be, will line up fully with the extension-determining response-independent underlying nature, whatever it may be. There are two kinds of possible mismatch. The first is that there may be instances of the underlying nature that do not produce the characteristic specified responses. Wright (as reported by Gundersen 2006) gives the example of a strange but tiny sample of H_2O somewhere in Siberia that does not produce the usual water-y sensory appearances in human beings, even in the most favourable of conditions. Surely, given its chemical structure, we would still want to class the Siberian sample as a case of water. The second kind of mismatch is the 'fool's gold' phenomenon: there may be a substance that would consistently generate the characteristic specified responses associated with a certain concept while realizing such a different underlying nature to that discovered in the overwhelming majority of the reference-fixing cases that we would not want to include it within the extension of the concept. The contingency in force means that neither of these mismatches can be ruled out a priori.

That last point clears the way to seeing why, from one perspective, this can all look like a bug in the reference-fixing model. Recall that the canonical way to formulate response-dependence is in terms of a principle such as: it is a priori true that x is F if and only if O judges x to be F in C. But the contingency in force within the reference-fixing model, as illustrated by the mismatches just described, means that that model plausibly loses access to the a priori aspect of this principle. Indeed, notice that even if we formulate the key principle only as a sufficient condition (which, in effect, is what Proudfoot does), fool's-gold-style cases will still do enough, within the reference-fixing framework, to disrupt its a priori status. Still, one person's bug is another person's feature, and, as I am about to argue, when it comes to the application of response-dependence to the concept of intelligence, the gap in question is a positive boon. So, although the fact that Proudfoot doesn't lay emphasis on the a priori aspect of the usual reference-fixing principle means that that dimension of the view has not so far been a focus of the analysis presented in this paper, it turns out that it matters.

Here's why. When the reference-fixing model is applied to the concept of intelligence, and played out by way of the Turing Test, here's how things go: the intelligent responses made by the appropriate interrogator in the test provisionally determine the extension of the concept of intelligence to include the computer, but the account of what is finally sufficient for the computer to instantiate the property of intelligence is given in terms of realizing the computational profile that constitutes the underlying nature of intelligence. Here there is no threat of circularity, and what fundamentally makes that so is the gap (the contingency of the connection) between the responses that provisionally fix the referent of the concept and the extension-determining underlying nature. So, given that, on my interpretation, Turing thinks of intelligence on the reference-fixing model, he is, on this score anyway, headache-free. Proudfoot's Turing, who endorses the transparency model, isn't.

The second advantage of the reference-fixing version of the response-dependence view of intelligence also emerges in the vicinity of the point that, on that model, a particular inner computational profile could constitute the underlying essence that ultimately determines the extension of the concept of intelligence. If we place the Turing Test within this framework, we arguably strengthen its resilience in the face of a common objection. That objection turns on the idea that we can conceive of situations in which, although the machine passes the test, we would nevertheless withhold the attribution of intelligence, on the grounds that it features an inappropriate inner profile. The behaviouristic version of the Turing Test is often thought to succumb to this very objection, precisely because it cannot make room for any such withholding. And now notice that a Turing Test placed within the transparency-based response-dependence framework faces an analogous difficulty, since, within that framework, there is nothing more to a machine being intelligent than its being judged as such by the average interrogator in the imitation game. Although this carries us beyond behaviourism, it doesn't create enough conceptual space for the inner computational profile of the machine to count as evidence regarding the presence of intelligence. By contrast, the inductive version of the Turing Test would create adequate conceptual space, but, of course, like the behaviourist version, it will struggle to do proper justice to the interrogator-centred, as opposed to task-centred, form

of the test, as Turing himself develops it. However, with the test restaged within the reference-fixing framework, things look encouragingly different. When the interrogator judges that a machine is intelligent, what she does is provisionally determine the extension of the concept of intelligence to include that machine, by presuming that it instantiates the relevant computational profile. It is entirely consistent with this approach that we may later withdraw the attribution of intelligence, if we discover that, in fact, no such instantiation obtains. And that plausibly renders the Turing Test resistant to the tabled objection, and so enables it to take on a newly resilient status within AI.

8 Concluding Remarks

Reading historical thinkers through the lenses of contemporary philosophical frameworks is always a hazardous business, but it can also be illuminating. Turing's claim that intelligence is an 'emotional concept' is rightly interpreted as indicating that he held a response-dependence view of thought and intelligence. But response-dependence comes in two varieties, and Turing is better characterized as pursuing the reference-fixing variety, rather than the transparency one. That fact, I have argued, makes his view all the more compelling, in part because of what it tells us about the form and resilience of the Turing Test.

Acknowledgements Many thanks to Crispin Wright for advice on how to distinguish between the two main models of response-dependence, and to Adrian Haddock for making me think about the circularity issue. Any mistakes that remain in this material are down to me. Thanks also to my anonymous referees for helpful comments that enabled me to improve the paper.

Funding None.

Compliance with Ethical Standards

Conflicts of interest The author declares no conflict of interests.

References

Braithwaite, R. B., Jefferson, G., Newman, M., & Turing, A. M. (1952) *Can automatic calculating machines be said to think?* in (Copeland 2004b), 410–32.

Copeland, B. J. (2000). The Turing Test. *Minds and Machines, 10,* 519–539.
Copeland, B. J. (2004a), *Introduction to 'Can automatic calculating machines be said to think?*', in Copeland 2004b, 487–93.
Copeland, B. J. (Ed.). (2004b). *The essential Turing: The ideas that gave birth to the computer age*. Oxford: OUP.
Dennett, D. C. (1981), True believers: The intentional strategy and why it works, in A.F. Heath (ed.) *Scientific explanation*, Oxford: OUP. Reprinted in J. Haugeland (ed,) *Mind design II: philosophy, psychology, artificial intelligence*, Cambridge: Mass: MIT Press, 57–79, from which page numbers are taken.
Fodor, J. (1987). *Why there still has to be a language of thought his Psychosemantics: The problem of meaning in the philosophy of mind* (pp. 135–167). Cambridge: MIT Press.
Genova, J. (1994). Turing's sexual guessing game. *Social Epistemology, 8,* 313–326.
Gundersen, E. B. (2006). *Making sense of response-dependence*. St Andrews: PhD thesis. University of St Andrews.
Johnston, M. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society, Supplementary, 63,* 139–174.
Moor, J. (1976). An analysis of Turing's Test. *Philosophical Studies, 30,* 249–257.
Moor, J. (2001). The status and future of the Turing Test. *Minds and Machines, 11,* 77–93.
Pettit, P. (1991). Realism and response-dependence. *Mind, 100*(4), 587–626.
Piccinini, G. (2000). Turing's rules for the imitation game. *Minds and Machines, 10,* 573–585.
Proudfoot, D. (2013). Rethinking Turing's test. *Journal of Philosophy, 110*(7), 391–411.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3,* 417–424.
Traiger, S. (2000). Making the right identification in the Turing Test. *Minds and Machines, 10,* 561–572.
Turing, A. M. (1948). Intelligent machinery: a report. *National Physical Laboratory*, in (Copeland 2004b), 410–32.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59, 433–460. Reprinted in (Copeland 2004b), 441–464. Page numbers in the text refer to the reprinted version.
Wright, C. (1988). Moral values, projection and secondary qualities. *Proceedings of the Aristotelian Society, Supplementary, 62,* 1–26.
Wright, C. (1989). Wittgenstein's rule-following considerations and the central project of theoretical linguistics. In A. George (Ed.), *Reflections on Chomsky* (pp. 233–264). Oxford: Basil Blackwell.