Minds & Machines (2012) 22:191–212
DOI 10.1007/s11023-012-9272-4

The Externalist Foundations of a Truly Total Turing Test

Paul Schweizer

Received: 22 September 2011 / Accepted: 22 February 2012 / Published online: 8 March 2012

Abstract The paper begins by examining the original Turing Test (2T) and Searle's antithetical Chinese Room Argument, which is intended to refute the 2T in particular, as well as *any* formal or abstract procedural theory of the mind in general. In the ensuing dispute between Searle and his own critics, I argue that Searle's 'internalist' strategy is unable to deflect Dennett's combined robotic-systems reply and the allied Total Turing Test (3T). Many would hold that the 3T marks the culmination of the dialectic and, in principle, constitutes a fully adequate empirical standard for judging that an artifact is intelligent on a par with human beings. However, the paper carries the debate forward by arguing that the sociolinguistic factors highlighted in externalist views in the philosophy of language indicate the need for a fundamental shift in perspective in a Truly Total Turing Test (4T). It's not enough to focus on Dennett's individual robot viewed as a system; instead, we need to focus on an ongoing *system of* such artifacts. Hence a 4T should evaluate the general *category* of cognitive organization under investigation, rather than the performance of single specimens. From this comprehensive standpoint, the question is not whether an individual instance could simulate intelligent behavior within the context of a pre-existing sociolinguistic culture developed by the *human* cognitive type. Instead the key issue is whether the artificial cognitive type *itself* is capable of producing a comparable sociolinguistic medium.

Keywords Artificial intelligence · Chinese room argument · Computational theory of mind · Mental content · Semantic externalism · Turing tests

P. Schweizer (✉)
Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh,
Edinburgh EH8 9AD, UK
e-mail: paul@inf.ed.ac.uk

The Turing Test and 'Strong' AI

What would be required for a computational artifact to count as genuinely intelligent in a manner comparable to human beings? Turing (1950) famously proposed an answer to this question, and the controversy launched by his position is still underway. Turing replaced his opening question 'Can (or could) a machine think?' with the more precise and empirically tractable question 'Can (or could) a machine pass a certain type of test?', where the test criteria are framed in terms of *behavior* that is typically held to signify intelligence in the case of human beings. In particular, the original 'Turing Test' (2T) is based entirely on *linguistic* inputs and outputs. Linguistic performance is an apt choice as a pivotal criterion of intelligence, since human language is perhaps our most distinctive feature as cognitive agents, and is an essential medium through which most of our high level mental achievements are developed and expressed. Hence human language will retain a central role throughout the ensuing discussion.

In brief, (the standardized version of) Turing's test is an 'imitation game' involving three players: a computational artifact and two humans. One of the humans is the 'judge' and can pose questions to the remaining two players, where the goal of the game is for the questioner to determine which of the two respondents is the computer. If, after a set amount of time, the questioner guesses correctly, then the machine loses the game, and if the questioner is wrong then the machine wins. Turing claimed, as a basic theoretical point, that any machine that could win the game a suitable number of times has passed the test and should be judged to be intelligent, in the sense that its behavioral performance has been demonstrated to be indistinguishable from that of a human being.

Historically, there has been disagreement regarding the proper interpretation of Turing's position. Some have claimed that the 2T is proposed as an operational *definition* of intelligence (e.g. Block 1981, French 2000), and as such it has immediate and fundamental faults. However, in the current discussion I will adopt a weaker reading and interpret the test as comprising an empirically specifiable criterion for when intelligence can be legitimately *ascribed* to an artifact. On this reading, the main role of behavior is inductive or evidential rather than constitutive, and so behavioral tests for intelligence do not provide a necessary condition nor a reductive definition. At most, all that is warranted is a *positive* ascription of intelligence, *if* the test is adequate *and* the system passes. In the case of Turing's 1950 proposal, presumably the test is deemed adequate in terms of parity of input/output performance with human beings, and hence purports to employ the same operational standards that we tacitly adopt when ascribing intelligence to our fellow creatures.

Turing's 1950 paper touches upon many fascinating topics and possibilities, the majority of which will not be considered in the following discussion. Instead, primary focus will be placed on the 'standardized' 2T as just delineated, in conjunction with selected developments and refinements of his seminal insight relating computation to mentality. Turing's original discussion places emphasis on the notion of a computational 'thinking machine' able to perform successfully in the verbal imitation game, while McCarthy's subsequent (1955) proposal, which introduced the term ‘Artificial Intelligence’, is based on “the conjecture that every aspect of learning or any other feature of intelligence” can in principle be simulated by a computational artifact. Somewhat later, Newel and Simon (1976) made the much more explicit and powerful claim that “the necessary and sufficient condition for a physical system to exhibit general intelligent action is that it be a physical symbol system”, i.e. a machine that produces and manipulates a changing collection of symbol structures.

In line with this more comprehensive and explicit variation on the theme, Turing’s original insight, which could potentially be construed in terms of a behavioristic engineering project, has now been transformed into a basic position in a revolutionary new *science* of intelligence. According to the widely embraced ‘computational paradigm’, which underpins cognitive science, Strong AI and various allied positions in the philosophy of mind, computation (of one sort or another) is held to provide the scientific key to explaining mentality in general and, ultimately, to reproducing it artificially. The paradigm maintains that cognitive processes are essentially computational processes, and hence that intelligence in the physical world arises when a material system implements the appropriate kind of computational formalism. So this broadly Computational Theory of Mind (CTM) holds that the mental states, properties and contents sustained by human beings are fundamentally computational in nature, and that computation, at least in principle, opens the possibility of creating artificial minds with comparable states, properties and contents.

Searle’s Critique

Probably the most high profile criticism of the 2T in particular, along with strong AI and the CTM in general, is provided by Searle (1980), where he puts forward his celebrated Chinese Room Argument (CRA). Although the structure of his original thought experiment may seem deceptively simple and straightforward, both the overall target and the underlying polemical strategy of his argument are perhaps surprisingly less than transparent.[1] The CRA is based on a hypothetical scenario in which Searle, a native speaker of English who knows no Chinese, is locked in a room with a massive rule-book written in English. Given Turing’s 1950 exposition of digital computers, this is a very apt rendition of serving as a human implementation of a formal procedure. Searle receives Chinese inputs on bits of paper and mechanically follows the instruction manual to produce outputs in Chinese script. For the sake of argument, we are asked to suppose that the manual is so good that he is able to fool native speakers and pass a Chinese 2T.

But Searle doesn’t understand Chinese, and doesn’t even know basic Chinese vocabulary. He’s just mechanically transforming arbitrary symbols according to a program of rules, while the inputs and outputs are, to him, totally meaningless. Hence Searle takes the CRA to refute the view that success at Turing’s imitation game constitutes an adequate standard for the ascription of authentic intelligence,

[1] See Harnad (2002) for a concise discussion/analysis of the CRA.

understanding or mental states, because in the hypothetical scenario he has passed a Chinese 2T while understanding nothing of Chinese. He also concludes that implementing a program with the requisite syntactical input/output profiles is theoretically inadequate as a criterion of intelligence. In direct opposition to Newell and Simon, Searle maintains that computation is neither a necessary nor a sufficient condition for mentality, and so the general project of CTM and strong AI are summarily rejected, along with the Turing test.

However, I would contend that the original CRA serves (at most) to refute only the 2T in particular, and not CTM or Strong AI in general, since acceptance of a computational theory of mind does not entail acceptance of the 2T. Many of Turing's successors in the fields of cognitive science and artificial intelligence would endorse a broad version of the computational paradigm, holding that mentality in the physical universe is to be explained via the realization of the right type of abstract formal procedure, without accepting the 2T as providing a sufficient condition for ascribing mentality or real intelligence to an artifact. The two themes are clearly separable, and it is possible to embrace a computational approach to the mind without accepting Turing's original and quite minimalistic standard (this basic point is made, e.g. in Rey 2002). As will be explored later in the paper, one can advocate much more stringent criteria for the success of Strong AI than the mere verbal imitation game proposed by Turing.

Moreover, from the comparatively narrow conclusion of the original CRA, viz., that successfully implementing a conjectured program for processing Chinese syntax does not constitute a sufficient condition for understanding Chinese, Searle makes a very brisk transition to the universally quantified assertion that there is no program or abstract formal procedure *whatever*, the implementation of which would be sufficient for true intelligence. Presumably, Searle's tacit supposition is that the CRA supplies a general recipe or template which can be readopted to fit and refute *any* given variation on the computationalist theme, and hence the basic CRA strategy is powerful enough to establish his universal conclusion. But I will shortly argue that this supposition is incorrect.

The Systems Reply

In the style of Turing's 1950 article, Searle considers and dismisses a number of anticipated objections to his view, the first of which he dubs the 'systems reply'. A defender of the computational theory of mind might argue that perhaps Searle in isolation doesn't understand Chinese, but that's not the point, because the whole system that produces the behavior—room plus manual plus Searle—does understand Chinese. Searle has a two pronged response to this objection. One prong is the 'homuncular' claim that *he* is the only locus of understanding in the room, and if he doesn't understand Chinese, then nothing else about the system does. As a bald *assertion* this certainly has a good deal of common sense appeal, although it clearly leaves something to be desired as an *argument* against those who would deny its truth. Searle's second prong is an 'internalization' tack: suppose Searle were gifted with a photographic memory, and could memorize the rule book. Then the entire set-up could be internalized, and Searle could perform the rule governed manipulations simply by consulting his memory, sitting outside under a tree. Searle himself would then *be* the whole system, but he still wouldn't understand Chinese.

At this stage I will withhold comment regarding the logical status of Searle's 'internalization' strategy, but agree with his *conclusion* that the systems reply is unconvincing. In order to reach this conclusion, I would prefer to endorse Putnam's (1981) critique of the 2T and reapply it to the systems defense. Putnam argues that passing Turing's original test is not a sufficient condition for concluding that the computer genuinely understands or refers to anything with the strings of symbols it produces, because the computer doesn't have the right sort of relations and interactions with the objects and states of affairs *in the real world* that its words are supposed to be about. To illustrate the point; if the computer has no eyes, no hands, no mouth, and has never seen or eaten anything, then it is not talking about hamburgers when its program generates the string of English symbols 'h-a-m-b-u-r-g-e-r-s'—it's merely operating inside a closed loop of syntax. In terms of the original CRA scenario, the systems reply is inadequate because the system in question is incapable of *doing* the right sorts of things.

In sharp contrast, *our* talk of hamburgers is intimately connected to *nonverbal* transactions with the objects of reference. There are 'language entry rules' taking us from nonverbal stimuli to appropriate linguistic behaviors. When given the visual stimulus of being presented with a pizza, a taco and a kebab, we can produce the salient utterance "Those particular foodstuffs are not hamburgers". And there are 'language exit rules' taking us from linguistic expressions to appropriate nonverbal actions. For example, we can follow complex verbal instructions and produce the indicated patterns of behavior, such as finding the nearest Burger King on the basis of a description of its location in spoken English. Mastery of both of these types of rules is essential for deeming that a human agent understands natural language and is using expressions in a correct and referential manner—and the hapless 2T computer lacks *both*, as does the entire CRA set-up when viewed as a system.

The Total Turing Test

Hence the standard 2T is fundamentally inadequate as a test for *understanding*, because the range of behavior it takes into account is far too limited. It relies solely on verbal input/output patterns, and these alone are not sufficient to evince a correct *interpretation* of the manipulated strings. Language is primarily about *extra-linguistic* entities and states of affairs, and there is nothing in a cleverly designed program for pure syntax manipulation which allows it to break free of this closed loop of symbols and demonstrate a proper correlation between word and object. As Dennett (1980) observed in his original peer commentary on Searle's argument, such 'bed-ridden' programs are far too weak to underwrite any positive inferences. When it comes to judging human language users in normal contexts, we rely on a far richer domain of evidence. And this critically undermines the notion that the standard 2T is justified on the basis that it relies on evidential parity with our practices in ascribing mentalistic predicates to our fellow creatures.[2]

Even when the primary focus of investigation is language proficiency and comprehension, sheer *linguistic* input/output data is not enough. The inherent limitations of mere conversational performance naturally suggest a strengthening of the 2T, later named the Total Turing Test (3T) by Harnad (1991), wherein the repertoire of relevant behavior is expanded to include the full range of intelligent human activities. This will require that the computational procedures respond to and control not simply a teletype system for written inputs and outputs, but rather a well crafted artificial body. Thus in the 3T the scrutinized artifact is a *robot*, and the data to be tested coincide with the full spectrum of behaviors of which human beings are normally capable. In order to succeed, the 3T candidate must be able to do, in the real world of objects and people, everything that intelligent people can do. So perhaps the 3T will appear to enjoy empirical symmetry with the human case and hence constitute a sufficient condition for attributing mentalistic predicates to artifacts? Harnad (and a great many others) certainly think so. Thus Harnad expresses a widely held view when he claims that the 3T is "…no less (nor more) exacting a test of having a mind than the means we already use with one another… there is no stronger test, short of *being* the candidate" (p. 49). And, of course, the latter state of affairs is not an empirical option.

At this stage, the incorporation of behavior in the actual world is motivated simply in terms of providing a more adequate standard of evidence for true understanding. When humans talk about hamburgers, we normally assume that they would be able to *recognize* one upon presentation, be able to *distinguish* a hamburger from a cheese omelette, etc. If this turned out not to be the case, then we would have good reason to doubt that they really understand the meaning of the word. However, this is not yet to invoke causal interactions with objects and substances in the environment in a particular theory of the semantics for natural language per se, which will occur in section "Semantic Externalism".

The Robot Reply

As it happens, the 3T is already anticipated by Searle in his 1980 article, and in his 'robot reply' to the CRA he dismisses even this highly elevated criterion. As before, the first prong of Searle's rejoinder is 'homuncular'; he augments the original CRA by supposing that he is still locked in the room, but now some of the Chinese characters he receives are codes for digitalized inputs from the robot's sensory transducers, and some of the output symbols now control the motors inside the robot's body and make it move its arms and legs. Even so, all Searle is doing (perhaps remotely, from inside a control room), is manipulating uninterpreted

[2] Shieber (2007) provides a valiant and intriguing rehabilitation/defense of the 2T, but it nonetheless remains a 'bed-ridden' standard that neglects crucial behavioral data, such as mastery of salient language exit and entry rules. Ultimately Shieber's rehabilitation in terms of interactive proof requires acceptance of the notion that *conversational* input/response patters alone are sufficient, which premise I would deny for the reasons given. The program is still operating within a closed syntactic bubble.

syntax. He has no idea what is going on outside the control room and the manipulated syntax still has no intentional content. Searle cannot *see* the greasy hamburger that the robot's photographic sensing apparatus has transduced into Chinese code, nor is he *trying to grasp it* by outputting the relevant effector code controlling the robotic hand.

This is no doubt a plausible rendition of Searle's benighted predicament *qua* control room homunculus, but it invites the question as to why this should constitute a pertinent concern. And the question has now become much more pressing because, unlike the original 2T, a vast number of the pivotal inputs and outputs in this more demanding case are no longer directly present to Searle. In order to pass this combined linguistic and robotic test, the artifact must perform physical behaviors in accordance with the language entry and exit rules appropriate for a correct grasp of Chinese. Hence the relevant input/output boundaries for the *system* under appraisal extend far beyond Searle-the-homunculus. The robot's sensing devices will comprise relevant input boundaries, while its artificial body and limbs will constitute the salient output interface for manifesting the scrutinized behavior. Thus, a great many of the crucial inputs and outputs for the robot undergoing the 3T are no longer comprised by the symbols with which Searle has acquaintance. Instead, the formal expressions manipulated by Searle, as an executive subunit, are fed in from and into various other modules of the robot's processing structure and are not themselves the inputs and outputs which form the basis of the test, and which constitute the evidence for the intended conclusion that the *robot* understands Chinese.

Contrary to Searle's response to the anticipated robot reply, the conclusion of the 3T is *not* that a mere control room homunculus would understand Chinese. So if we apply the systems approach to the entire artifact that passes the 3T, as Dennett explicitly suggests in his original peer commentary, then Searle's preliminary rejoinder to the robot reply appears to be a non sequitur.[3]

The Robot as a System

The 3T robotic-system is fundamentally disanalogous with the initial 2T/CRA scenario, yet Searle tries to address Dennett's sharpened objection by applying the same internalization strategy as before. In his author's response to Dennett's commentary, as well as in a later (1994) reply to Harnad, Searle attempts to break free of his homuncular confines by claiming that he could in principle realize the entire *robotic system* himself and still not understand Chinese. It's far from clear that this claim expresses a cogent theoretical possibility, so I will not belabor assorted variations on the fanciful set-up and try to adjudicate. The methodological value of the thought experiment has transgressed its boundaries and I will simply grant Searle's far-fetched hypothesis for the sake of argument. If Searle could

[3] Alternatively, Rapaport (2006) argues that human neuron firings are also just a form of uninterpreted syntax, so that what the homunculus Searle in the control room is doing is no different from what our brains do. And if our brains understand natural language, then there's no reason to deny this of Searle in the room, at least not just because all he's doing is manipulating uninterpreted syntax.

perhaps manage, in some arcane and convoluted manner, to realize the entire system and *become* the robot following its instructions, then we will concede that Searle himself, considered as a conscious subject, would not understand Chinese.

Nevertheless, this still isn't enough to establish the desired conclusion, since it fails to demonstrate that the overall *system* doesn't understand Chinese. The attempted inference trades on an equivocation: 'Searle' the narrow center of conscious subjectivity is quite distinct from 'Searle' the body/brain complex viewed as an objective and highly multifaceted arrangement of matter and energy. It is 'Searle' under this latter description that implements the robot, while it is 'Searle' under the former description who does not understand Chinese. Yet there are plainly many aspects, properties and attributes of 'Searle' the complex dynamical system which are not applicable to 'Searle' the conscious subject. For example: Searle the conscious subject is not himself able to follow instructions in spoken Chinese and behave in the indicated manner (because, granting Searle his own point, he doesn't understand any Chinese). Hence, if given the instruction in Chinese to raise his right hand and hop on his left foot, Searle the conscious subject does not possess the ability to process the instruction and perform the appropriate action. However, Searle-the-complex-system *can* process the instruction and perform the indicated behavior, as evidenced by the robot's ability to pass the salient Turing test. So from the fact that Searle the homunculus does not possess a given property or attribute it does not follow that Searle-the-system does not. Hence the form of Searle's proffered counterargument is logically invalid and his conclusion does not follow from the premises.

I would therefore contend that Searle has failed to provide a successful counterexample to the claim that passing the 3T is a sufficient condition for attributing intelligence or understanding to the overall system under scrutiny, and hence that Searle's introspective considerations are not able to deflect Dennett's combined robotic-systems reply. So at this point in the debate ultimately stemming from Turing's 1950 proposal, the CRA-type strategy can make no further progress. Searle's internalist approach serves, at most, to refute the adequacy of the original 2T, and cannot be generalized to the vastly strengthened sensorimotor version of the test. Hence the CRA alone does not provide a template which can be readopted to fit and refute any given variation on the computationalist theme. Searle's general anti-computationalist conclusion requires additional theoretical machinery for its defense.[4]

Of course, the failure of a purported counterexample does not in itself imply that the positive claim endorsed by Searle's opposition has been confirmed.[5] But given the highly elevated standards of the new test, along with the inadequacy of Searle's attempted rebuttal, the onus would now seem to lie in the other court. Indeed, many would hold that the 3T marks the culmination of the dialectic and, at least in principle, constitutes a fully adequate empirical standard for judging that an artifact possesses genuine mentality.

[4] For example, in (1984) and (1990) Searle makes some of the background machinery more explicit.

[5] This fact, in the context of the 2T, is also noted in, e.g. Copeland (2001).

However, I will now attempt to advance the debate by moving in a direction away from the restrictive internalism of the CRA. Two fundamental tenets underlying Searle's view are that (a) minds have mental contents (semantics), while (b) computational syntax, by itself, is neither constitutive of nor sufficient for semantics. The real driving force behind the CRA is to serve as a polemical tool in support of (b). The trouble is that in the 3T we are no longer dealing with computational syntax, *by itself*. Instead, we are dealing with a fully functioning robot able to *behave* in a manner indistinguishable from a human being, and hence which seems to evince a correct *interpretation* of the manipulated strings. In rejecting the sufficiency of the original 2T, I've agreed with Dennett and Putnam that mastery of the appropriate language entry and language exit rules is a necessary condition for an adequate test of artificial intelligence. But is mere behavioral success on its own sufficient?

In the remainder of the paper I will retain Searle's insistence on meaning as essential to mentality, as expressed in his fundamental tenet (a) above, but replace his 'psychologistic' version of semantics with the less traditional *externalist* view. This shift in semantic theory yields some interesting implications both for AI and for a strengthened Turing test. I will try to move the debate forward by exploring some key features of externalism which I think cast serious doubt on the idea that a behaviorally successful 3T robot understands language in a manner at all comparable with the paradigmatic human case, and that the expressions generated by the computational artifact are genuinely referential. Finally, I will argue that the *sociolinguistic* factors highlighted in externalist views indicate the need for a fundamental shift in perspective. It's not enough to focus on Dennett's individual 3T robot viewed as a system; instead, a Truly Total Turing Test (4T) needs to focus on an ongoing *system of* such artifacts. Hence my overall conclusion will be that the 3T is still too weak, and that a truly comprehensive test should evaluate the general *category* of cognitive organization under investigation, rather than the performance of single specimens.

Semantic Externalism

Externalist views in the theory of meaning and reference first put forward by Kripke (1972), Putnam (1975) and Burge (1979) highlight essential features of natural language (NL) semantics not present in the case of the 3T artifact as currently depicted. The conclusion of Putnam's highly influential Twin Earth Argument (TEA) is that the internal cognitive states of individual language users radically underdetermine linguistic meaning—generally, there's nothing 'in the head' strong enough to fix reference for terms in natural language. According to this view, no mere internal configuration of a cognitive system, be it computational, neurophysiological or conscious/phenomenal, is able to capture the intended objects of linguistic reference. Hence the representational capacities of internal states are, in the general case, too weak to support the referential burdens of natural language.

On Putnam's account, the naive 'psychologistic' approach is fatally flawed because it ignores two essential aspects of meaning and reference. One (1) is the role of direct causal interaction with the environment when language is acquired and used: natural kind terms such as ‘water’, ‘aluminum’, ‘gold’, ‘tiger’, etc., make indexical appeal to actual specimens or paradigm cases *in the world*—so causal relations via perception, demonstrative pointing and utterance production in the intersubjectively accessible public domain determine what these words actually refer to. There is no internal encoding or representational state sustained by the individual agent that is powerful enough to do this.

Second (2), the traditional internalist approach ignores what Putnam calls the ‘division of linguistic labor’, epitomized by the reliance on *experts* who set the standards for the entire linguistic community and underwrite the reference relation in cases where relevant microstructures and/or objective membership conditions *are* known. It is by *acquiring* a natural language within a particular sociolinguistic community and using it within this shared framework that we are able to refer successfully. For example, the average English speaker can use the word ‘gold’ to talk about the actual substance, even though they may not know the periodic table, may not know that gold is the element with atomic number 79, and probably don’t know in practice how to distinguish real gold from, say, chalcopyrite. Most people have had causal/perceptual interactions with samples of the metal itself, and thereby have direct indexical access to the substance the word names. But the precise technical details of the extension of ‘gold’ are uncovered by relevant experts in the field, and it is upon their expertise that our linguistic practice implicitly depends, and not upon our own internal representations, concepts or psychological states. ‘Gold’ patently does *not* mean ‘any arbitrary material with a set of phenomenal characteristics such that my subjective concepts or representational capacities are not able to distinguish it from the element with atomic number 79’.

And such externalist criteria cannot be felicitously detached from, nor be rejected as irrelevant to, an *individual’s* NL understanding. If someone, say Arnold, were to claim that the above disjunction *is* what he means, then Arnold would be open to the charge that when he uses the word ‘gold’ in that manner he is no longer speaking English. Furthermore, these externalist criteria are also basic to linguistic *communication*, since, e.g. two agents cannot communicate if they are not even talking about the same things when using the same words. Hence if I am unaware of the non-standard aspects of Arnold’s ideolect, then he and I fail to communicate whenever either of us uses the word ‘gold’. And clearly these observations about ‘gold’ generalize across the entire English language.

Kripke first argued for externalist factors in the case of proper names, where the referent is not determined by a definite description or ‘cognitive content’ entertained by the language user, but rather is historically anchored in a brute association between name and individual, and where a causal chain of social practice preserves the correlation established through this ‘initial baptism’. As above, Putnam extends the analysis to include natural kind terms such as ‘gold’ and ‘water’, while Burge extends it even further to include a host of *conventional* taxonomic kind terms like ‘arthritis’, ‘sofa’, ‘contract’, ‘brisket’, etc. The semantic reliance on factors outside the bounds of the individual language user thus appears to be an indispensable and ubiquitous feature of NL. In short, language is a communal, historically evolved phenomenon, where the meaning of words is not determined by individual representations or internal states, but is a public, external matter, determined by objective microstructural regularities, causal chains, relevant experts and accepted practices in one's sociolinguistic clan. Putnam concludes that we must give up the view that meanings are concepts or mental entities of any kind. According to his famous adage, "Slice the pie any way you like, meanings just ain't in the head."

Robotic Reference

But if meanings ain't in the heads of individual human agents, then they're certainly not in the data bases of computational artifacts. So, in light of (1) above, if the 3T robot's natural language capabilities are simply installed as part of its overall program, then it will not have the necessary history of causal interactions with the objects of reference, and its symbolic activities will remain semantically ungrounded. On the foregoing widely accepted model of 'direct' reference, there is an essential causal and chronological link that semantically tethers an individual's linguistic behavior to its environmental context. The relation of reference is founded on a history of causal interactions between the agent and the entities and states of affairs in the world that it uses language to talk about, where salient aspects of the external world have *causally impinged* upon the agent. The word 'water' as used by typical human beings is intimately linked to a long history of associations based on experiences of seeing, drinking, washing with, and being immersed in various samples of environmental H_2O, where these experiences are all *caused* by the liquid itself, giving the agent direct indexical access to water, as the word was acquired and integrated into its overall linguistic framework.

At the moment it's obviously rather difficult to envision exactly how a robot might be designed to pass the 3T. But *if* the computational core of its abilities were simply implanted via some sophisticated natural language processing (NLP) software, in combination with vast data bases, world models, etc., then the concomitant lack of an historical chain of interaction with the real world poses a serious theoretical question regarding the semantic import of its linguistic outputs. When a token of the term 'water' is emitted by the robot, all shiny and fresh off the assembly line, how could it possibly *mean* 'the liquid with the same underlying microstructure as the stuff in the environment that I've interacted with when I acquired the word'? If the robot has not yet had any physical interactions with actual water, then it would seem to be semantically no better off than the original 2T computer, encased in a closed loop of syntax.

However, the issue becomes more subtle. Turing's original test has been deemed inadequate because it fails to incorporate vital behavioral data in terms of the language entry and exit rules basic to *demonstrating* a veridical grasp of NL, and this has motivated the transition to the robotic 3T currently at issue. This is an operational consideration based on the behavioral evidence required to judge whether or not a given system can be said to understand a particular language, and as such it is not tied to any more detailed theory of NL semantics per se. In contrast, externalism constitutes a very specific position in the philosophical analysis of meaning and reference, and one which places vital emphasis on causal interactions between language user and world. So in this sense, the two strands may begin to superimpose. In order to pass the Total Turing Test, the robot must also *behave* in all the appropriate manners with its artificial body, and after it's been around for awhile it will have acquired its own personal history of causal interactions with water—and then the theoretical waters will become muddied. It seems clear that when the robot is taken fresh off the assembly line its linguistic outputs will lack a requisite external referent, as in the original 2T scenario. But after prolonged bodily and verbal actions performed in the real world, as demanded by the 3T in order to demonstrate a genuine grasp of English, it seems that a case could perhaps then be made for proper semantical grounding.

So I do not present the issue of (1) as an insurmountable obstacle nor a conclusive, in principle objection, but rather as an interesting and potentially important case of dissimilarity with the semantic analysis of naturally occurring cognitive systems. Indeed, it might turn out to be impossible to construct an artifact capable of passing the 3T without first *training* it to use language in the real world and thereby providing the requisite history of causal interactions.[6] So the issues raised by (1) are perhaps not insuperable. However, I think that the sociolinguistic aspects invoked in the division of linguistic labor in (2) above underpin a much more serious difficulty when evaluating the robot, and one which, even if it could possibly be overcome in the case of an individual artifact, nevertheless suggests that this would still not be enough to attain full parity with the overall performance capacities of human cognitive architecture. Hence the ramifications of factor (2) will then serve to motivate the claim that even the combined linguistic/robotic 3T is intrinsically too limited, and that a conceptual shift in goal posts is required for a Truly Total Turing Test. But first factor (2) itself will be explored in more detail.

The Sociolinguistic Community

In line with the foregoing observations regarding the central role of linguistic culture, in order for the robot's linguistic activities to be genuinely referential, the robot would have to *acquire* its linguistic fluency through interaction not just with the environment as required by factor (1), but as a member of the relevant sociolinguistic community. And again, this is very different than having its language processing abilities simply implanted as a formal program, particularly if this program were predesigned in terms of some chosen external target language(s).

If the robot did not *learn* its language via extended participation with an actual and embodied linguistic culture, within a shared physical and social context, then it will not be a valid member of any such community, and consequently it will be unable to rely upon the historical chains of name use, division of linguistic labor and

[6] Interestingly, Turing considers the possibility that the best way to produce a machine able to pass the 2T might be to "follow the normal teaching of a child". However, when describing the 'child programme' he observes that "It will not be possible to apply exactly the same teaching process to the machine as to a normal child. It will not, for instance, be provided with legs …" indicating that Turing, at this point, is speculating about a learning program, not a genuine robot (although he does subsequently conjecture about engineering enhancements, which seem to anticipate the robotic 3T).

other cultural practices central to our referential success. Putnam gives the analogy that natural language is not like a hammer, a tool that can be wielded successfully by an individual. Instead, language is a cooperative social venture, more like operating a steam ship or perhaps a large industry. As *bone fide* members of the English speaking 'linguistic cooperative', we're automatically plugged into this ancient and highly structured communication system, a living cognitive network through which we inherit and access the meaning of our words.

For the linguistic activities of single human beings to be semantically grounded, the individuals must belong to and participate in such a communication network, a network that is anchored to a continuous presence extended in real time and space. People first have direct causal interactions with various persons, places, objects and natural kinds in their immediate surroundings, and by learning and exercising their linguistic behaviors in this shared environment, they enjoy direct indexical access to the referents of the corresponding terms. But via full membership in this same NL community, they also gain linguistic access to people, places, objects, substances and states of affairs *remote* in both time and space. I've never been to Madagascar, and Isaac Newton died long before I was born. Nonetheless, through membership in the English speaking NL sociolinguistic coop, I'm plugged into this ongoing, far reaching and exceedingly powerful communication network, and am able to use English words to successfully *talk about* Isaac Newton and Madagascar, even though I've been in direct personal contact with neither.

However, *if* the 3T robot's English capabilities are simply installed as part of some highly sophisticated NLP software package, *then* it will lack the essential history of having acquired these abilities through interaction with and participation in an actual, embodied community. Its 'semantics' will be purely internal and solipsistic, generated from files stored in its data bases and various coded representations supplied by its designers. And as Putnam's TEA convincingly shows, such internal states and structures are incapable of determining the reference relation for even such basic natural kind terms as 'water'.

So the issue at hand does not concern the bare *mechanics* of how one might design and build a computational artifact with the behavioral capacities to pass the 3T. Granted, such an engineering project will dwell on the occurrent and real-time abilities, structural properties and causal powers of the physical device. Instead, the issue is one of subsequent *evaluation* of the robot with respect to its semantic and mentalistic properties, such as genuine intelligence, understanding, reference for its linguistic outputs, and the attribution of associated mental states and contents such as *believing that* snow is white, *knowing that* water is H_2O, *wanting to* pass the 3T, etc. And it is here that externalism, historicism and the sociolinguistic medium play a crucial role in both our conceptual framework and our actual practices. And, as will be argued below, it is also here that testing of the artificial *type* rather than just a token artifact is the salient level of analysis.

Of course, in the same vein as noted above, the combined linguistic/robotic standards of the 3T would require the robot to have extended dealings with human beings while it was undergoing the test, and one might then argue that after it had been around for some time and had sufficient verbal and other behavioral interchanges with humans, it would itself gradually *become* a card carrying member of the English speaking sociolinguistic coop, with full rights and privileges.[7] And while a case could perhaps be made that a successful 3T robot, fully integrated into human society, might eventually be deemed a legitimate member of the English speaking community, I will argue in sections "The Non-individualism of the Mental" and "A Truly Total Turing Test" that the issue nonetheless points to a fundamental feature of human mentality that has been entirely neglected in the standard test scenarios, and which this form of mere *integration* would fail to address.

But just to summarize the discussion so far; it's not a question of whether or not the robot can *behave* in a given sociolinguistic context in an appropriate manner and thus pass the combined linguistic and robotic 3T—we assume for the sake of argument that it *can*. Instead, the question concerns the conceptual adequacy of the 3T itself, and the main issue has been whether or not the robot's use of language is genuinely meaningful. Do its words *refer*, is it employing natural language to *talk about* extralinguistic objects and states of affairs, or is it merely producing syntactic strings as output in response to various surface stimuli in conjunction with cleverly designed internal models and formal recipes for symbol manipulation? Put in terms of the classic use/mention distinction in logic; is the artifact simply *mentioning* linguistic expressions, 'displaying' items of formal syntax with no semantic content, or is it *using* these expressions in a robust and referential fashion? According to externalism, the latter question cannot be answered in the affirmative unless both conditions (1) and (2) above are satisfied. I do not argue that they cannot be satisfied, but merely that success at the 3T does not guarantee this. And in the following sections I will maintain that, even if they can be satisfied in the case of particular instances, this is still not enough for a truly comprehensive test.

The Non-individualism of the Mental

As stated above, the aim is *not* to explain and predict the behavior of the robot viewed as an ingenious piece of mechanical engineering. If the 3T is being used to test Strong AI's ultimate goal of creating a robot with an artificial *mind*, then this will require the successful application of our standard framework of *mentalistic* explanation to characterize and predict the robot's performance. In viewing a system *qua* mind, rather than just another complex physical or biological device, a crucial move is to apply the Belief–Desire (BD) framework of explanation paradigmatic of mental systems, wherein cognitive agents are seen as possessing a vast store of propositional attitudes, which rationally combine via psychological processing to *cause* actions.

[7] This is perhaps comparable to a situation where unsuspecting earthlings crash land their space ship on Twin Earth. On day one they will still mean H_2O when they utter the term 'water', since that's the native interpretation of their language. But after they've lived on Twin Earth for sometime and had sufficiently many interactions with environmental XYZ, and been integrated into their new sociolinguistic clan, they will enter a grey area, and it is plausible to hold that they will eventually become grounded in Twin Earth semantics and mean XYZ when they say 'water'.

The well known basic scheme is to ascribe to the agent assorted beliefs and desires, where beliefs depict the way the world is and how things work, while desires supply goals—possible future states of the world that the agent wants to become actual. Then we explain/predict that (other things being equal) the system acts to achieve its desires in light of its beliefs. This constitutes what it means to be a rational mental agent. For example, if Mary walked to the bar because she *wanted* a shot of whisky and she *believed* that she could obtain one there, then that's a perfectly full and complete account of her action. In terms of standard psychological explanation, the foregoing account is not in need of, nor is it improved by, the addition of further details concerning the neurophysiological substrate of Mary's beliefs, the biomechanics of limb movement, the psychophysical correlations between the reception of electromagnetic radiation and Mary's visual experiences that enable her to find the bar, etc. According to the canonical BD framework, rational actions are caused by propositional attitude states in virtue of the representational *content* of these states. And if, at the *mental level* of explanation, such actions are caused by states individuated in terms of their content, then appeal to these content laden states, rather than their mechanical underpinnings, is the salient mode of explanation and prediction.

And it is here that the sociolinguistic dimension of semantic externalism begins to seriously impinge upon the more traditional preconceptions underlying both the 2T and 3T. As Burge perspicuously observes, the mental attributes and contents ascribed to individual agents depend in an essential manner on the practices and conventions of one's external sociolinguistic community. He supplies a number of counterfactual illustrations where an individual's mental contents differ while their entire physical and internal mental histories, viewed in isolation from their social context, remain the same. The differences in content stem from differences outside the individual "considered as an isolated physical organism, causal mechanism or seat of consciousness." In turn, such differences in content are normally taken to indicate differences in mental states and events. Burge's perhaps startling conclusion: various *mental* states and events are not fully determined by what's going on in an *individual's head*. Instead, they rely in an inextricable manner on the encompassing sociolinguistic milieu.

From this it follows that *human* mentality is essentially non-individualistic—it depends crucially upon a sociolinguistic context that transcends personal boundaries. The particular contents and hence the very identity of the mental events and propositional attitude states that characterize an agent are not wholly determined by what's going on internally. Instead, the identity of these states and events is inextricably dependent on the surrounding sociolinguistic medium. Similarly, this sociolinguistic medium is not a product of any human individual, but rather is the legacy of the human cognitive *type*. And this indicates that the phenomenon of mentality is, in a very fundamental sense, sustained at the type rather than the token level. Tokens of the human cognitive variety have the mental states and contents they do in virtue of their dependency upon the capabilities of the human cognitive type, where the type is responsible for the sociolinguistic medium in which the tokens are embedded.

And in turn, this indicates that tests for *artificial* intelligence and mentality that focus merely on the performance of single specimens are guilty of ignoring a fundamental aspect of the phenomenon in question. If the test is one for a genuine mind that has been artificially engendered, then the real standard should concern the capabilities of the synthetic cognitive *type* itself rather than token artifacts. And the issue becomes particularly acute given that the 3T simply presuppose human sociolinguistic culture as a background condition, as a starting point to which the artificial agent can adapt. But because of the non-individualism of the mental, the successful 3T robot is then, at best, not a case of purely *artificial* intelligence, but rather a curious hybrid: an artifact assimilating to a pre-existing background context produced and sustained by an alien cognitive kind, and where this alien *human* context makes an essential contribution to the attributed mental states and contents of the robot. The resulting form of 'mentality' is then a cognitive mongrel, a blend of both human and artificial elements.

Thus a fundamental defect of the 3T, when construed as a test for a truly artificial mind, stems from the fact that the ascription and the very identity of the robot's mental states, events and contents is inextricably dependent on its sociolinguistic medium, which in this case has been produced by an entirely different and *non-artificial* cognitive architecture. So when testing a purported case of genuinely artificial intelligence, the pertinent question becomes—is the artificial *type* capable of producing and sustaining this essential sociolinguistic medium? If not, then the ascription of mental states, events and contents to successful 3T tokens of this type is inescapably derivative, and the overall cognitive kind to which such tokens belong is not itself capable of supporting mentality in a manner equivalent to the capabilities of the human cognitive category. In terms of a Truly Total Turing Test (4T), the artificial type must be capable of autonomously generating the essential sociolinguistic medium, and not simply adapting to what the human kind has already produced.

So, following Burge, and contrary to the more naive traditional conception, basic mental phenomena such as propositional attitude states are *not individualistic*. But because the 2T and 3T are concerned only with artifact tokens, these tests are framed in purely individualistic terms, and tacitly assume from the outset a conception of mental contents and states which externalist considerations undermine. And it is salient to note that externalism itself is much more compatible with the basic 'operational' methodology of Turing tests, because it does not appeal to Searle's internalist, homuncular features of an agent. Instead of relying on first person introspection, externalism is much more dependent on observation of the system *from the outside*, incorporating causal, behavioral, social and historical factors. And this reflects our actual practice in ascribing mentalistic states and properties to our fellow creatures. Our practice is based upon such factors as observed in the intersubjectively accessible domain, and does not make essential appeal to what might be going on inside someone else's head. And once the fundamental importance of the sociolinguistic medium is recognized, as brought into sharp focus by externalist considerations, then the individualistic standards built into the assorted versions of the Turing test as so far conceived must be replaced by operational considerations at the non-individualistic *type* level. Hence the deep-seated contribution of the sociolinguistic milieu highlights the need to shift the locus of scrutiny from an artifact token to the *artificial cognitive type*.

A Truly Total Turing Test

As argued above, semantic externalism provides some very strong considerations in support of the move from the token to the type level in a fully adequate test of artificial mentality. But the mere fact that both the 2T and 3T start out by presupposing *human* natural language as a background condition already indicates the need for this fundamental shift. Unless the type of cognitive architecture under scrutiny has the independent capacity to generate and sustain the kind of sociolinguistic context assumed as a starting point, then the test is still too weak. As an illustration of the basic conceptual need to test the overall category of cognitive architecture and not just the performance of individual specimens in a given and pre-existing NL context, consider the following hypothetical scenario that does not in any way depend on semantic externalism. Suppose, purely for the sake of argument, that Fodor and Pylyshyn (1988) are right, and that human cognitive processing takes place via an underlying symbolic Language of Thought (LOT). And suppose that an artificial neural network is developed which is so cunningly designed that it can be trained to exploit the external symbol system of *human* natural language, extract information and produce the requisite input/output patterns enabling it to pass the 3T. But imagine further (and still in line with Fodor and Pylyshyn) that this type of connectionist architecture, on its own, is totally incapable of generating the productive and systematic external symbol system of human NL, and the allied sociolinguistic culture, upon which its 3T success depends. Hence as a cognitive processing *type* it would fail the 4T, and its token success at the 3T would be wholly dependent upon exploiting a system of exterior cultural scaffolding which it is intrinsically unable to produce. In such a case, the 3T success of individual artifacts is clearly parasitic upon the pre-existence of the more advanced linguistic capabilities of the human cognitive type, and the natural conclusion to draw is that the artificial neural network *does not* possess capabilities fully equivalent to the human symbolic LOT architecture.

As an indigenous phenomenon, human language use and linguistically characterized mental content depends on membership in a sociolinguistic community, and furthermore, one that has been created and is sustained by *conspecifics*: the successful linguistic and cognitive activities of human individuals are inseparable from immersion in an historically evolved culture of intelligence, where this culture is itself the product of *human* cognitive processing. So with the move to recognizing the fundamental importance of linguistic culture in intelligent human behavior goes a concomitant shift in emphasis from tokens to the level of the general cognitive kind. The capacities at the general type level are conceptually prior to the relevant features and capabilities of individual specimens.

And this highlights a fundamental deficiency in the approach employed by Turing-type tests as envisioned so far. In light of the foregoing, it is not theoretically sufficient for a comprehensive test of *artificial* intelligence to focus merely on the performance of individual artifacts. Instead, such a test should focus on the overall capabilities of the general cognitive category to which these tokens belong. So the manner in which the robotic 3T is designed still reveals a crucial disanalogy with the human case. Not only can individual human beings exhibit the salient patterns of verbal input/output behavior required by the original Turing Test, and full blown mastery of the language entry and exit rules required by the combined linguistic and robotic Total Turing Test, but in addition it was the *human* cognitive type, to which such individual specimens belong, that has produced natural language and this advanced culture of intelligence in the first place. And it is with other tokens of this *same* type that we are inter-twinned as a sociolinguistic community, and upon whom our referential success, as well as the very identity of our linguistically attributed mental states, co-depends.

In stark contrast, the computational artifact involved in the 3T is *not* a member of this same type. It has an alien and artificial cognitive structure which is quite possibly incapable, at the type level, of ever producing natural language or the kind of sociolinguistic context which is a necessary background condition. Indeed, both NL itself as an advanced structural phenomenon, and the associated cultural and communicative network which engender it, are simply *presupposed* as a starting point for the 3T robot. And this indicates that a conceptual shift is required in order to frame a truly thorough test of artificial intelligence along the operational lines originally proposed by Turing. When it comes to rigorously evaluating the capabilities of an artificial cognitive architecture, it is not enough to scrutinize the performance of individual specimens. The question of genuine intelligence must advert to the type rather than the token level, since the performance of an isolated agent presupposes a social context which transcends the individual. As noted in section "The Sociolinguistic Community" above, after many years of behavioral interaction and social integration, a token of an alien cognitive type, such as an individual 3T robot, *might* become a naturalized member of some native NL community, and hence use human language in a semantically grounded manner. But the robot's referential success would then be parasitic upon a communal and linguistic framework produced by an entirely different kind of cognitive architecture, as would the very contents of its propositional attitude states used to explain and predict its behavior. Hence the salient capabilities of its *own* cognitive type would remain untested. What is required in a fully comprehensive examination of artificial mentality is evidence that a community of the robot's *own conspecific* is capable of producing and sustaining the kind of sociolinguistic medium that the 2T and 3T merely presuppose.

So on these grounds Dennett's robotic, *individual* system reply to the CRA is still too weak to establish his positive view. The scrutinized 3T artifact is simply planted in a pre-existing natural language community, within a social context of intelligence produced by a radically different cognitive type—namely the same type as its designers! Because advanced human NL is assumed from the start, the behavioral capabilities of the robot can presuppose sophisticated, pre-structured linguistic inputs for free, and these can serve as triggers for appropriately complex responses. The cognitive architecture can thus be tailor made to allow success in a highly developed and specialized context, a context about which its designers have exhaustive foreknowledge, and one which quite possibly could never have been generated by the type of architecture in question. Thus we need to evaluate not only the capabilities of the artifact viewed as a system (as opposed to a mere control room homunculus), but also the capabilities of an ongoing *system of* such artifacts, to determine whether or not the type of cognitive architecture in question is able to produce the social medium of intelligence that is simply taken for granted as a background condition in the 3T.

The cognitive processes that underpin the use of natural language in human begins are also the cognitive processes which created these languages to begin with. The initial inputs were the basic environmental stimuli available to prelinguistic creatures, and *systems of tokens* of the human cognitive type transformed these inputs, via a cooperative, interactive and incremental development, into the extremely rich and subtle communication network currently sustained by human NL communities. Thus the successful 3T behavior of individuals artifacts, within a pre-existing sociolinguistic context (e.g. Twentieth to Twenty-first century English speaking civilization) is not a sufficiently rigorous test of the general type of cognitive structure underlying these individual performances. Instead, the real capabilities of the structure-type can only be manifested by starting at square one. So, from the point of view of comparative *total* capabilities the 3T is still inadequate, since the artifact is supplied with a prefabricated stage on which to perform acts of post hoc imitation. And this sort of 'potted' test could conceivably be passed by a well designed computational puppet, rather than by an instance of a robust and genuinely intelligent brand of cognitive organization (see my earlier discussion in Schweizer (1998) for allied points motivated by a different set of considerations).

Incidentally, I think this explains why the standard science fiction scenario of an advanced alien life form, regardless of its chemical composition or internal processing structure, is always a more convincing hypothetical case of true intelligence than a 3T artifact. In contrast to a robot, the alien life form will have evolved its own sociolinguistic culture of native intelligence, in response to its primitive environmental stimuli, rather than simply exhibiting programmed capabilities *simulating* real intelligence in a pre-existing context for which it was tailor made by its designers. In this speculative case of an advanced alien life form, say Martians for convenience, the *type* of cognitive architecture in question would already have passed a 4T, as has the human race. Hence there is nothing species-centric about the 4T, and the intellectually and linguistically advanced Martians could pass the 4T independently of any association with particular *human* NL communities.

Furthermore, the fact that the Martian race has already passed the 4T supplies a natural basis for then ascribing real intelligence and linguistic understanding to the *individual* Martians we might encounter. By acquiring their language via participation in a genuine, albeit alien, sociolinguistic medium, and causal interaction with their own local physical environment, individual Martian language use would thereby be semantically grounded. Hence a Martian who studied English as a second language on Mars would presumably be able to master and *understand* a great deal of English, by *translating* this foreign language into its own semantically grounded 'native tongue'. Most likely there would be sufficiently many differences between Mars and Earth (and Martian and English) that there would be many things that it would *not* be able to grasp without first flying to Earth for a visit. But there is no reason to suppose that, once it had spent some time here and was sufficiently acquainted with our practices that it could not pass an extraterrestrial version of the 3T. And the appropriate conclusion to draw in such a case would be that the individual Martian possesses an authentic understanding of English.[8]

Conclusion

There is a profound difference between the historically manifested capabilities of the human cognitive type, as opposed to the standards incorporated in behavioral tests that merely scrutinize the ability of secluded artifacts to imitate the forms of behavior that we have already developed and sustained in response to much more basic environmental stimuli. And this crucial disparity in standards obtains even if tokens of the artificial type were to do as well or even better on the 2T or the 3T than individual humans. For example, even though we may take skilful chess playing in the case of human beings as a sign of intelligence, we can nonetheless maintain that a chess playing program is not 'genuinely intelligent', irrespective of how well it plays chess. Skilful human chess players are members of the same cognitive kind that developed the overall social practice of game-playing, originated the game of chess itself, and then designed the program that now beats us at it. In contrast, the type of computational procedure to which the program belongs is capable of *none* of these prior achievements, even though it now wins the game. Similarly, it's quite conceivable that single instances of Dennett's robotic system could pass the 3T (because tailor made for this task), and yet the underlying cognitive architecture could fail miserably at the 4T. And again, the natural conclusion to draw would be that the type of system in question *does not* possess intelligence in a manner fully comparable with human beings. Instead, its behavioral successes are still essentially derivative, and parasitical upon the capacities and attainments of a fundamentally dissimilar cognitive type.

So this points to a significant shift in conceptual perspective: a truly total test should focus on the capacities of the cognitive category as a whole, rather than on the performance of isolated tokens. The original 2T simply presupposes human NL and its associated cultural underpinnings as a precondition for the purely verbal imitation game. Although a vast improvement in many ways, the combined

[8] A similar but less extreme point holds in regard to human members of different native NL groups. Since we are all member of the same cognitive type the 4T is not an issue. So a French person who diligently studied English as a second language in Paris and then came to London would presumably be able to understand enough English to pass an English 3T. And this is because the French person is a member of the French sociolinguistic community, and hence is already semantically grounded in a human NL, and can thereby understand English by first translating it into French. So one might be able to learn a language 'purely syntactically', but only through the pre-existence of a semantical foundation in some prior *interpreted* language. And clearly this is asymmetrical with the case of a newly assembled 3T robot. I would like to thank an anonymous reviewer for bringing this (and the Martian learning English case) to my attention as potential objections to my view.

linguistic and robotic 3T still incorporates this intrinsic limitation, by again focussing on an artifact token, specifically designed to mimic the full range of intelligent human behavior within a cultural network produced by a different category of cognitive agent altogether. But rather than consider the imitation of *our* intelligent behavior by specially designed tokens, the criterion for a 4T should be whether an ongoing system of these artificial agents is capable of producing a comparable sociolinguistic medium of intelligence, starting from the same primitive environmental inputs as our pre-linguistic forebears.[9]

In summary, my main thesis is that even if a computational agent could pass the 3T, this would not be sufficient for the attribution of genuine mental states and contents, unless it were semantically grounded in an actual sociolinguistic community and physical environment. Since it's conceivable that a cunningly designed robot might pass the 3T *without* this grounding, it follows that merely passing the 3T is not sufficient for the attribution of genuine mental states and contents. Furthermore, even if a robot passed the 3T and were semantically grounded (say, by *learning* English in the actual world and as part of an English speaking community), this would still not be a sufficient test for artificial intelligence on a par with humans, since the sociolinguistic medium presupposed by the 3T was produced by the human and not the artificial cognitive type. It's conceivable that tokens of the artificial cognitive type could pass the 3T *and* be semantically grounded, while the artificial type itself is intrinsically incapable of producing and sustaining the essential and pre-existing sociolinguistic medium of intelligence to which it is able to adapt. Hence the need for a 4T to test the performance capacities at the type level. To count as a *bona fide* case of *artificial mentality*, the robotic cognitive type would have to demonstrate the capacity to produce its own sociolinguistic medium of intelligence, just like an advanced alien life form or the human race.

References

Block, N. (1981). Psychologism and behaviorism. *Philosophical Review, 90*, 5–43.

Burge, T. (1979). Individualism and the mental. In P. French, T. Euhling, & H. Wettstein (Eds.), *Studies in epistemology, vol. 4, midwest studies in philosophy* (Vol. 4). Minneapolis: University of Minnesota Press.

Copeland, B. J. (2001). The Turing test. *Minds and Machines, 10*, 519–539.

Dennett, D. (1980). The milk of human intentionality. *Behavioral and Brain Sciences, 3*, 428–430.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1–2), 3–71.

[9] As with the 3T, the proposed test framework is quite futuristic (as even the original 2T is now turning out to have been), since the paper is not concerned with the practicalities of carrying out actual assessments, but rather with the operational standards which, *in principle*, are required to attain parity with the full range of data available in the human case. The evidence has taken tens of thousands of years to manifest itself, and we would have to somehow collapse the timeframe in order to apply the same standards to an artificial cognitive type. One possibility would be to use computer modelling to run 'evolutionary' scenarios in much faster than real time, where simulations of communities of the artificial agents could perhaps yield answers about long term 4T capabilities.

French, R. (2000). The Turing test: The first 50 years. *Trends in Cognitive Sciences, 4*, 115–122.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines, 1*, 43–54.

Harnad, S. (2002). Minds, machines and Searle 2: What's wrong and right bout Searle's Chinese room argument? In J. Preston & M. Bishop (Eds.), *Views into the Chinese room: New essays on Searle and artificial intelligence* (pp. 294–307). Oxford: Oxford University Press.

Kripke, S. (1972). *Naming and necessity*. Harvard: Harvard University Press.

McCarthy, J. (1955). *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. http://www.formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

Newel, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery, 19*, 113–126.

Putnam, H. (1975). The meaning of 'meaning'. In *Mind, language and reality*, Cambridge: Cambridge University Press.

Putnam, H. (1981). Brains in a vat. In *Reason, truth and history*, pp. 1–21, Cambridge: Cambridge University Press.

Rapaport, W. J. (2006). How Helen Keller used syntactic semantics to escape from a Chinese room. *Minds and Machines, 16*, 381–436.

Rey, G. (2002). Searle's misunderstanding of functionalism and strong AI. In J. Preston & M. Bishop (Eds.), *Views into the Chinese room: New essays on Searle and artificial intelligence* (pp. 201–225). Oxford: Oxford University Press.

Schweizer, P. (1998). The truly total Turing test. *Minds and Machines, 8*, 263–272.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*, 417–424.

Searle, J. (1984). *Minds, brains and science*. Harvard: Harvard University Press.

Searle, J. (1990). Consciousness, explanatory inversion and cognitive science. *Behavioral and Brain Sciences, 13*, 585–596.

Searle, J. (1994). The failures of computationalism. *Think, 2*, 68–71.

Shieber, S. (2007). The Turing test as interactive proof. *Nous, 41*, 33–60.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.