In Defence of a Reciprocal Turing Test

Fintan Mallory¹

Received: 5 November 2019 / Accepted: 16 November 2020 / Published online: 7 December 2020
© The Author(s) 2020

Abstract

The traditional Turing test appeals to an interrogator's judgement to determine whether or not their interlocutor is an intelligent agent. This paper argues that this kind of asymmetric experimental set-up is inappropriate for tracking a property such as intelligence because intelligence is grounded in part by symmetric relations of recognition between agents. In place, it proposes a reciprocal test which takes into account the judgments of both interrogators and competitors to determine if an agent is intelligent. This form of social interaction better tracks both the evolution of natural intelligence and how the concept of intelligence is actually used within our society. This new test is defended against the criticisms that a proof of intelligence requires a demonstration of self-consciousness and that semantic externalism entails that a non-embodied Turing test is inadequate.

Keywords Turing test · Intelligence · Recognition · Externalism · Artificial intelligence

The Turing test is taken to provide an indication of several things; whether a machine 'thinks', is 'intelligent' or possesses 'a mind'. Turing never made much of an effort to distinguish between the different mental properties we might be interested in ascribing to machines and, in any case, these properties are typically taken to come as a package. To be intelligent *is* to be able to think, to have a mind, to be a concept-user, and so forth. Even if we take possession of these properties to be a gradient affair, it remains plausible that they are dependent on each other. In the following, I will use the word 'intelligence' as a catch-all term for these features.

Critics of the test point out that intelligence also involves a myriad of abilities such as spatial reasoning, problem-solving, planning, self-awareness and so on, which the test does not directly engage with. If the test is supposed to provide an 'operational definition' of intelligence, it only seems to cover a small part of what intelligence really is. And if by 'intelligence', we simply mean 'that property the

---

✉ Fintan Mallory
fintan.d.mallory@kcl.ac.uk

¹ King's College London, Strand, London WC2R 2LS, UK

imitation game tests for', then there's little reason to think we've made progress when a machine has passed. This is the first problem which any discussion of the Turing test must address.

Turing was aware that some people might take his test to simply change the subject.¹ He acknowledged that the word 'intelligence' is equivocal and that there are many possible ways it could be precisified. Aside from the usual confusions that arise from calling products of intelligence 'intelligent' (e.g. moves in chess, books, computer programmes), we quite reasonably ascribe intelligence across the animal kingdom and beyond. There is certainly a sense in which my cat is intelligent and another sense in which he isn't. The same is the case for ants, ant colonies, and ant-eaters. Turing's writings suggest that he took arguments over which of these senses was the one correct sense to be futile. If he was interested in describing 'thinking' and 'intelligence' in ways that applied to octopuses and other intelligent life forms, then he presumably wouldn't have specified a complex dialogical task as a means of identifying it. After all, he could have stipulated that intelligence amounted to the ability to solve problems of a certain level of computational complexity. All this would achieve is to detach his discussion of intelligence from the context of the lively debates public debates about artificial intelligence in which he was working.

I take it that what makes the prospect of artificial intelligence interesting and controversial is not the idea that we can have more efficient calculators or more accurate problem solvers but that there may be agents very much like ourselves yet with radically different cognitive architectures and historical origins. It may be possible to imagine superhuman forms of intelligence or intelligences that are profoundly different from our own (though this latter claim can be disputed) but the kind of intelligence we are interested in is the intelligence required for a machine to think the kinds of thoughts we think and use the concepts we use. This idea will be at the heart of the following discussion. When we ask if machines can be 'intelligent' or can 'think', I take the interesting question to be, can this machine have the kind of mental states a human has? This is something that a cat may never achieve but that machine might. In other words, the imitation game doesn't test for mindedness but for like-mindedness.² The interrogator attempts to determine whether the contestant is a fellow human being, not whether they are an intelligent agent (in the way that an octopus might be). For the rest of this paper, the word 'intelligence' should be understood to mean human-like intelligence.

We need to be clear about how we view the function of the concept of intelligence, how it is typically wielded, and what we would achieve in applying it to

¹ 'You might call it a test to see whether the machine thinks, but it would be better to avoid begging the question, and say that the machines that pass are (let's say) 'Grade A machines' (Turing 1952, p. 495). There is another 'everyday notion' of intelligence which is context-sensitive and according to which we might say that someone is very intelligent for a toddler or very intelligent for a cat. That is not of interest here.

² Turing is quite open that the intelligence he is interested in involves 'simulating the behaviour of the human mind'. While this focus has been criticized, notably by Hayes and Ford 1995, I am assuming that this goal is a more interesting project than the development of better measurement devices or mathematical problem-solvers.

machines. Discussions of the Turing test often make different assumptions about what kind of concept intelligence is; whether it's content can be reduced to an operation definition, whether it is response-dependent, an empty metaphor, or tracks a real property. Some of the test's strongest critics defend a form of semantic externalism which has direct implications for not only for the possibility of artificial intelligence but the design of any test for it. The assumptions we make about the concept of intelligence have consequences for the design of any Turing-style test as the test relies on an interrogator's grasp of the concept of intelligence and their intuitions about what this involves. We cannot discuss the test without considering the nature of the concept that the judge is using. A central idea of this paper is that, when we set up a Turing-style test, we are engaged in an act of triangulation. We implicitly endorse the judge's grasp of the concept of intelligence (just as we assume that our thermometer is working when we use it to tell the temperature). The judge is not taken to be an expert but they are assumed to have an adequate understanding of the concept corresponding to the property in which we are interested. The notion of intelligence that is captured by the test is, inasmuch as there is one, the 'everyday' notion we apply to other humans. I assume that people grasp and apply this concept when engaging with each other and do so with little explicit thought; unaware of the necessary and sufficient conditions for intelligence but able to recognise it when they see it. It is a concept which serves to track other concept-users and it thereby distinguishes beings that reason like us from those that don't reason at all. This paper will argue that such a concept is fundamentally connected to our processes of determining whether we are engaging with other intelligent beings and that this is something which any serious test of intelligence must assess for.

I suggest that the idea that there is some merely implicit notion of intelligence that we use on a regular basis might solve a strange and under-discussed tension between two ideas in Turing's work. Turing didn't identify being *able to think* with being *taken to think*—he explicitly disavowed the idea that the philosophical problem of whether a machine can 'think' could be resolved by a Gallup poll. At the same time, he couched the conclusion of his famous *Mind* paper in a claim about how linguistic usage might shift over time. He does not say whether or not that usage would be measured correct relative to some objective standard of correctness, e.g. how things really are, but he does seem to take it to play an important role in our considerations. This issue will be picked up again at the very end of this paper.

The first part of this paper discusses the relationship between interpretations of the Turing test and how we think about the contents of our concepts. The section which follows argues that the traditional imitation game is not sufficient to test for the presence of intelligence. A novel form of the test is proposed which takes into account the importance of tracking intelligence in our concept of intelligence while the final section defends this proposal against some old and new objections.

We can't engage with the question of artificial intelligence unless we know something about the concept of intelligence. Different approaches to the Turing test assume different approaches to concept possession. While this paper will speak loosely of 'applying concepts' and will make some claims about the consequences of doing so, I don't wish to hook the interpretation which follows to a robust or

fine-grained theory of conceptual content. The reader is invited to translate any claims about concept application into whatever vocabulary they feel most comfortable with.

I don't profess that the main claim of this paper is entirely original, its antecedents will be made clear in the final section, and neither do I wish to claim that this is what Turing was always thinking, though I do believe it can be motivated by his text. I do think though there is an important kernel of truth in this idea that deserves a decent hearing and could help inform future research.

1 The Concept of Intelligence in Interpretations of the Turing Test

Interpretations of Turing's test can be distinguished by whether they understand the interrogator to be administering the test or whether they take the test to be administered by some third party. Behaviourist or 'operationalist' interpretations take the interrogator to be administering the test and thereby take the conditions which give rise to the interrogator's judgement that they are in dialogue with an intelligent human to provide the conditions of application for the concept of intelligence.³ According to this reading, Turing identifies intelligence with the circumstances that give rise to the judgment that an agent is human, in other words, a dispositional property of the agent. As Ned Block puts it, '[i]ntelligence (or more accurately, conversational intelligence) is the disposition to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be' (Block 1981, p. 11). Implicit in this interpretation is the idea that the content of a concept can be reduced to the observable phenomena it tracks.

Critics point out that this definition is not extensionally adequate. On the one hand, it over-generates; something could be disposed to act this way and still not be intelligent (e.g. Blockhead, Chinese rooms etc.). On the other hand, the definition can under-generate by failing to capture cases of genuine intelligence which cannot be measured this way (e.g. if the contestant is uncooperative). This objection is typically followed by the claim that there is some further property picked out by the concept of 'intelligence' which cannot be tested for by the imitation game; consciousness, creativity, a soul or something else. These critics, at least those that advocate what Block calls *psychologism*, argue that 'intelligence' denotes a deeper property internal to the agent which gives rise to these dispositional properties.

The alternative interpretation assumes that the test is administered by a third party who is merely using the interrogator as a measuring device. Accordingly, the

³ Hence Block's claim: 'The Turing Test conception of intelligence offers a necessary and sufficient condition of intelligence' (Block 1981, p. 14). Searle: 'The Turing test is typical of the tradition in being unashamedly behavioristic and operationalistic' (Searle 1980, p. 423). An especially influential behaviorist definition of intelligence was put forward by Turing' (Shannon and McCarthy 1956, pp. v–vi). See also French (2000). These days, it is generally accepted that the original test aims to provide a sufficient but not a necessary condition for being regarded as intelligent. Turing never indicates that he thinks that his test is the 'only' way to determine if a machine thinks. This fact does not preclude us from talking about machines 'passing' or 'failing' the test, one can still fail to meet a sufficient condition.

behaviour that is actually measured in the Turing test is not the activity or dispositions of the contestant but the behaviour of the interrogator responding to the contestant, specifically, their judgement of whether the contestant is human or not. An advantage this has over the behaviourist reading is that it takes into account the central position of the interrogator in the imitation game. It also importantly allows us to separate the significance of the concept applied from the behaviour that gives rise to the interrogator's judgement. The conceptual assumption here is that an analysis of the concept isn't reducible to its circumstances of application but must include some account of the interrogator's behaviour as well.

Saying that there must be something downstream from an application of a concept isn't particularly interesting unless we know something about what that is. Proudfoot's response-dependence interpretation is one way of doing this (Proudfoot 2013, 2020). According to the response-dependence interpretation of Turing's test: *x* is intelligent (or thinks) if, in an unrestricted computer-imitates-human game, *x* appears intelligent to an average interrogator.⁴

This reading holds that the purpose of the imitation game is to establish 'paradigmatic conditions for the application of the response-dependent concept of intelligence (or thinking)' (Proudfoot 2013, p. 239). The reason we need to specify paradigmatic conditions is to avoid the interference of atypical environmental conditions on the subject's response and to filter cases in which there would be interpersonal discrepancy among responses (Haukioja 2007; Pettit 1991). Just as the paradigmatic conditions in which we apply colour predicates will be under white (day-)light rather than under a red lamp, the imitation game gives us normal conditions under which an agent must appear intelligent to an average interrogator. Inasmuch as intelligence can be understood as a dispositional property, it is a disposition to bring about a response in another agent.

The error the behaviourist makes, according to this view, is to identify 'intelligence' with the conditions provoking a judge's response. From the response-dependence perspective, this is akin to identifying the concept red exclusively with the microphysical properties which give rise to the experience of redness. The response-dependence theorist, in contrast, holds that certain properties and concepts cannot be understood except in relation to the responses they bring about in agents (classic examples are colours and properties like nauseating, funny, and painful). The metaconceptual apparatus of response-dependence enables us to capture both these elements without reduction to either. It treats the claim that an agent is intelligent as more like the claim that 'water' is wet than 'water is H₂O'. An adequate analysis requires discussion of both the property which provokes a response and the response itself.

⁴ Note that this does not say 'iff'. The test embodies a claim about sufficiency rather than necessity. In contrast, the classic response-dependent account of 'red' tells us: *x* is red iff *x* looks red to normal subjects. The textual support for this interpretation comes from Turing's discussion of intelligence as an 'emotional concept' and his claim that, '[t]he 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' (Turing 1948, p. 431).

I have raised the response-dependence interpretation not because I think it is the most plausible reading of Turing's work (though I think it highlights several matters of importance which have been overlooked). The issue I wish to focus on is not what gives rise to the judgment that an agent is a human but what follows from it on the side of the observer. One doesn't need to endorse the response-dependence framework in order to claim that the significance of concepts isn't just determined by their circumstances of application (as is implied by the operational-definition reading) and the response-dependence account is one way among others of highlighting this. There are consequences to judging that something has a property and if we want to formulate a test for the property, we need to take these into account. When we reflect upon what is involved in judging that an agent is intelligent, we can get a better idea of what it takes for a machine to be intelligent.

Recall that the imitation game demands that an interrogator determine who is human in an exchange.⁵ There are times when it might be appropriate to regard 'humanity' as some kind of monadic property which an object might possess or not but there is obviously more to taking a person as a human than ascribing them a property. An analysis of what it is to be human that simply says that to be human is to bear the property 'is-a-human' is neither useful nor theoretically interesting. The virtue of the response-dependence theory is that it works in some sense to connect the question posed by Turing's test to the wider and deeper issue of what is involved in regarding another as a human.

Instead of defending the response-dependence view, which may be right or wrong, I want to make three—hopefully innocuous—claims. The first is that the relations underlying concepts or properties may be symmetric or non-symmetric. The second is that any test for a property is simultaneously a test for a concept (and vice versa). The third is that some concepts can be action-guiding. These ideas will become relevant in the discussion of Turing's test.

A core idea motivating some proponents of response-dependence theories is the idea that certain properties which have previously been understood as monadic are in fact grounded in relations and that one cannot truly grasp the concept corresponding to those properties unless one has a grasp of the subjective side of the relation. As a result, a test for such a concept or property is a test for the existence of such a relation, typically a relation between a stimulus and a response, i.e. the disposition of the stimulus to produce the response. The kind of test we use, though, must be sensitive to the properties of the underlying relation. For example, standard response-dependent properties such as 'red', 'funny', 'scary' or 'nauseating' are typically grounded by non-symmetric relations (or at least relations which can be non-symmetric). If something is red, it stands in the relation to an agent of reliably causing them to experience 'redness' (under normal conditions) and this informs how we would test if something is red. To test if something is scary, we might measure heart

⁵ Strictly speaking, the original gender-based form of the game asks the participant to identify whether the contestant is a man or a woman. The rest of this paper will focus on versions of the Standard Turing Test (STT) which ask a judge to determine if they are speaking with a machine or human (or 'intelligent agent').

rates and adrenaline; to test for funniness, we might measure whether the bearer makes people laugh. These tests are broadly non-symmetric in that they only test if the relation holds one way. A subject does not need to produce the sensation of ‘warmth’ in an object for that object to be considered warm.

Not all properties (or concepts of properties) can be understood in terms of non-symmetric relations. Whether or not someone is a ‘friend’ depends both on how you respond to them and how they respond to you (if you have a more pessimistic view of friendship, perhaps ‘ally’, ‘colleague’ or ‘lover’ might work better here). This doesn’t mean that you require direct access to their internal states but you do need to know that they have to judge you as their friend (or ally) for them to count as yours. If you think the concept of friendship can be asymmetric, you may not grasp the normal concept of friendship. In contrast to a non-symmetric test which involves presenting an agent with a possible instance of the property and asking them to judge whether it is present, a symmetric test requires that the stimulus–response relation works both ways.

This means that a test for a property is simultaneously a test for a concept since, when testing for the presence of a property, we must assume that the person whose perceptual/cognitive faculties are being used actually possesses the required capacities to identify the property. This means that we can understand our tests in two ways. Asking someone to look at a picture and judge whether it is blue or green might be a way of determining the colour of the picture or whether or not that person is colour blind. More generally, when we measure a physical property, we can simultaneously be said to be testing whether our measuring device is functioning.⁶ The difference between these tasks is a matter of emphasis rather than action. Ultimately, some degree of triangulation will be needed to determine that our measuring device or test-subjects are adequate (for example, we might introduce a statistical criterion). This triangulation is typically built into response-dependence theories by reference to a ‘typical observer’.

It’s worth mentioning that a person might possess the underlying capacities to identify a difference without possessing the explicit concept corresponding to this difference. Some languages lack terms for certain colours, but we may assume that the speakers of those languages still possess the primary capacity to discriminate between colours. I don’t want to take a strong stance on whether possession of a concept is possession of a word but one result of possessing a word for a property is that it allows us to make explicit our responses.⁷

The responses built into the concepts, ‘red’, ‘funny’, ‘nauseating’ and ‘morally obligatory’ are all very different. Typical response-dependent accounts involve some phenomenal or gustatory experience, perhaps an emotion or sensation, however, we don’t need to limit ourselves to these. Some responses will include the application of

⁶ This issue is discussed in depth by Turing and Wittgenstein in the latter’s *Lectures on the Foundations of Mathematics*.

⁷ Discussion of this issue in linguistics has largely centred around various forms of the Sapir-Whorf hypothesis (Regier and Kay 2009). Within philosophy, it has been connected to the rejection of an empirical ‘given’ and the idea that our direct perception of the world might be—in some sense—conceptually structured. Fortunately, the main claims of this paper are independent of this issue.

further concepts to the entity in question. If I respond to an object by judging it to be ‘red’, it’s plausible that my response also involves applying the concepts ‘coloured’ or ‘not-blue’ etc. These additional responses serve largely descriptive and classificatory purposes. As such, we can understand them as expressing the significance of the concept ‘red’. This is part of the down-stream significance of the concept which doesn’t fit directly into the response-dependence framework but which one might wish to capture.

While taking something to be ‘red’ may only guide those internal, cognitive actions which underpin your ability to represent the world, some of the most well-studied concepts are also *action-guiding*. For example, consider a concept like ‘morally obligatory’. We might try to think of this as a response-dependent concept (i.e. it depends upon agents capable of performing the action) but the statement that something is ‘morally obligatory’ iff it seems morally obligatory to a reasonable, decent agent, is missing an important feature. Something that is morally obligatory should also give you a *reason* to act a certain way. A person has not grasped the concept unless they have grasped this. If you think that something is morally obligatory but don’t also think that you have a reason to do it, then you may be regarded as confused. The right hand of the response-dependence conditional may not only include *seemings* but *demands* as well. It seems highly plausible that taking something to be ‘intelligent’ (or ‘thinking’) makes similar action-guiding demands on an agent. Calling something intelligent might mean that we need to adopt the ‘intentional stance’ to its behaviour or place its claims in the ‘space of reasons’. These are simply different ways in which one could articulate the idea of an ‘appropriate response’ to an intelligent stimulus.

The rest of this paper works from the assumption that the concepts ‘intelligent’ and ‘thinking’ are both *symmetric* and *action-guiding*. In other words, the concepts of ‘intelligence’ or ‘thought’ which are invoked in the Turing test cannot be fully analysed by an account of the conditions which lead someone to apply them. Furthermore, I will argue that the resources to develop this are contained within Turing’s paper by specifying some necessary properties of the ‘intelligence-recognition’ response.⁸

2 Response and Recognition

The imitation game as it is introduced by Turing is a non-symmetric test. By non-symmetric, I mean that the interrogator alone determines whether or not the contestants have passed or failed and that there is no corresponding risk and or pay-off for the interrogator.⁹ If the interrogator concludes that the machine is human, we

⁸ ‘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).

⁹ In the following, I will speak of ‘failing’ a Turing test as a shorthand for not being regarded as a human intelligence. This can happen to humans and machines alike as both are attempting to convince the interrogator that they are human. I do not mean to imply that the test provides a necessary (rather than sufficient) test for intelligence. Genuinely thoughtful, intelligent beings might still fail the modified test. But if they pass it, we should regard them as intelligent.

might loosely say that they have ‘passed’ the test, though it is important that we don’t thereby mistake passing the test as being a necessary condition for one’s being an intelligent agent. After all, if we are dealing with the traditional imitation game, if the machine has passed, the human confederate has ‘failed’.

My strategy in this section is to use some real-world examples to motivate the idea that intelligence must be at least a symmetric concept and probably transitive as well. As such, a reformulated Turing test should be symmetric. I take it for granted that most people aren’t convinced by the original non-symmetric test. We don’t need to appeal to imaginary cases like Blockhead in order to see this. Every day, people engage—often quite actively—with online chatbots under the misapprehension that they are speaking to intelligent, thinking beings. Yet it’s unlikely that, when shown the truth about their interlocutors, anyone would take the chatbots to be intelligent or engaged in thinking.¹⁰ One of the reasons we don’t take these machines to be intelligent is that we implicitly understand that it is not enough to dupe someone in order to be intelligent. Conversely, we typically think there is more to being intelligent than being taken to be intelligent. People have historically ascribed intelligence to celestial bodies and been incorrect to do so.

I believe the resources to address this are already contained within Turing’s paper and simply need to be teased out. To see this we need to notice a tension within the original formulation of the imitation game. Recall that rather than assessing a mixture of reasoning skills and problem-solving abilities, the standard fare of intelligence testing, the imitation game tests only for whether an agent is judged to be an intelligent agent or not (this is what determines whether an agent has passed the test or not). Any mix of skills might be sufficient to demonstrate intelligence without any individual skill being necessary. Turing appears to make no specific assumption about the necessary conditions for intelligence, except one. It is assumed that the interrogator is an intelligent agent and thus that an intelligent agent can serve as the interrogator for the imitation game. The interrogator must have the property for which they are testing.

In the case of some properties, as we develop means for identifying them; whether that is a scale, a chemical test, our process of determining if the property is present is slowly stripped of its human element. One standard is replaced with another. Whether or not something counts as water ceases to depend upon whether it tastes like water and becomes a matter of whether it contains H_2O . We are not in a position to do this in the case of intelligence. As measuring devices must be triangulated, any test for thought we formulate will ultimately be subject to *human* judgements concerning its accuracy. The response-dependence view would suggest that this arises because intelligence is ultimately a response-dependent concept. However, we don’t have to go that far to make the weaker claim that any method for determining if

¹⁰ Of course, arguing with a chat-bot is different from a well-conducted Turing test. The point here is merely that people generally don’t accept that a machine can think on the grounds that it can convince another person that it is a human. In other words, I don’t think that the traditional Turing test actually captures people’s intuitions about intelligence. In this sense, I agree with Block that something more is required.

something is intelligent is ultimately subject to human judgements of accuracy. If a machine has human-like intelligence, it should be able to do the same thing.

The idea that a core feature of being an intelligent, human agent is the ability to discriminate between beings with intelligence and those without it is not original (as anyone who has seen *Blade Runner* could tell you). Yet despite this, the asymmetric form of the test does not test for *the one thing which an intelligent agent is assumed to be capable of doing*, judging if another agent is intelligent. This is the tension which the Reciprocal Turing test hopes to redress. For our purposes *to recognise something as intelligent is to acknowledge its ability to serve as an interrogator in the imitation game*. An agent should not pass the Reciprocal Turing test if it could not serve as a judge in the Turing test. To serve as a judge, one must be able to reliably tell the difference between intelligent agents (those who could likewise be a judge) and ‘mere’ machines. This doesn’t mean that a judge in the Turing test must be an expert about machines or have a sophisticated theory of intelligence but rather that they are capable of applying the standard, everyday concept of intelligence which enables human beings to distinguish humanity from the rest of the world. The idea here is that Turing’s test assumes that the human interrogator who serves as a judge in the test can distinguish between human intelligences and non-human machines (and this is the only thing really assumed about them; no technical knowledge is required). I believe this gives us a decent reason to treat the ability to judge intelligence as an important part of intelligence itself.

To do this we’ll need a minimal account of what the judgment involved commits us to, what is down-stream from applying the concept. This is the observer-side part of the story. Unfortunately, Turing doesn’t provide much detail about what this judgement involves. In keeping with tradition, we’ll call this judgment *recognition*. One of the advantages of this term is that it captures that the judgment involved is not a spontaneous one-off act, an epiphany about which there is no going back, but rather interrogators may judge and reassess many times throughout the experiment.

We must be clear about this; the claim that a test for intelligence requires that we identify whether or not the agent can carry out a test for intelligence does not mean that such an ability is itself constitutive of intelligence.¹¹ We live in a world of online Blade Runners, Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHAS) which routinely, successfully distinguish human intelligence from machines. While these systems aren’t perfect, I doubt that anyone would regard even a perfect CAPTCHA that had a 100% success rate as intelligent. The ability to serve as a judge in the test is not itself a sufficient condition. I am not plunged into an existential crisis when a CAPTCHA fails to recognise me as an intelligent being. When this happens, I realise that the problem is with the

¹¹ This is where we break with Watt 1996’s Inverted Turing Test (ITT) (Watt 1996). While the ITT emphasises the important of mind-reading for intelligence, whether a system passes it depends solely upon its ability to judge others. The perfect captcha would be able to do this. The asymmetry discussed here is different to the ‘interpretive asymmetry’ emphasised by Collins, 1997, which concerned the way that a human interpreter could compensate for a computer’s mistakes better than a computer could compensate for a human’s.

CAPTCHA. Implicit in our understanding of intelligence is the idea that if an agent is not capable of recognising you as intelligent, it can't be very intelligent itself.

This suggests that for something to be intelligent, it is not merely sufficient that it is judged to be intelligent (as per the original Turing test) or that it can judge whether something is intelligent (as per Watt's 1996 Inverted Turing Test). When assessing if something is intelligent, we are also concerned with its ability to judge our intelligence; intelligence requires symmetry (symmetry is not the conjunction $R_{xy} \& R_{yx}$ but the conditional $R_{xy} \rightarrow R_{yx}$).

However, symmetry may be insufficient on its own. One reason that I don't regard my cat as overly intelligent is that he appears to regard the Roomba as intelligent (and possibly the washing machine). This is a judgment that I don't endorse and it does not reflect well on him. If an intelligence is something which can tell the difference between intelligences and non-intelligences, then to recognise an intelligence is to recognise something that is able to identify the difference between intelligences and non-intelligences. We only recognise something as intelligent if we think that it can reliably identify other intelligent beings. To track this, I only recognise a being as intelligent if I also recognise the beings it recognises as intelligent as intelligent. Again, this condition is already implicit within the experimental paradigm. When we select a person to serve as a judge for the test, we are implicitly trusting that their judgment is adequate. We are assuming that, if they are intelligent, then what they recognise as intelligent must be intelligent. This is the *transitivity* condition; $R_{xy} \& R_{yz} \rightarrow R_{xz}$.¹²

The core idea here is that intelligence is deeply connected with, at the very least, an implicit grasp of the concept of intelligence. This point is deeper than it may appear. When I speak of 'intelligence' here, I mean the concept that allows one to distinguish thinking things from non-thinking things. I think we should accept this ability as a minimal requirement for intelligence (in the relevant sense). To be a thinker, one must be able to distinguish thought from its contrary, non-thought. Non-thought can be understood in a variety of ways, from *res extensia* to one's environment. In any case, the claim here is that to be intelligent, one must have the ability to distinguish one's thought from the unintelligent word one inhabits. This brings us to the most frequently discussed property, reflexivity.

A standard objection raised against the Turing test is that self-consciousness is a necessary condition for thinking and the Turing test has nothing to say about this. This arises from an influential tradition in philosophy which holds that, when it comes to *res cogitans*, a *cogito* is non-negotiable. A test of an agent's conversational capacity doesn't tell us directly about their capacity for self-knowledge, awareness or transcendental apperception. To keep matters focused on the concept of intelligence, we might say that we would not recognise a machine as intelligent unless it could recognise itself as intelligent. Turing discusses this point in response to Geoffrey Jefferson's Lister Oration. Jefferson claims that it would not be enough for a machine to write a sonnet because, in order to be intelligent, the machine must

¹² The transitivity of any substantive recognition relation is elaborated with much greater depth in chapter nine of Robert Brandom's *A Spirit of Trust* (Brandom 2019).

‘not only write it but know that it had written it’ (Turing 1950a, p. 451). Turing’s response was effectively to accept that reflexivity cannot be tested but to reject it as a criterion for intelligence because accepting it would result in solipsism; ‘Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks’ (ibid).

This argument has been heavily criticised over the years and I won’t defend it here.¹³ It is sufficient for us to note that if a relation closed on a set is symmetric and transitive, it is also reflexive.¹⁴ We cannot coherently assume that an agent is capable of recognising us as intelligent and of recognising that those it recognises as intelligent can recognise agents as intelligent, without also recognising that it is itself intelligent. We will unpack this idea below but before we do so, it would be helpful to imagine what a Reciprocal Turing Test might look like.

3 The Reciprocal Turing Test

To recap, the relation the Turing test is trying to discover is symmetric but the traditional test is asymmetric. Testing for symmetry and transitivity requires different experimental parameters to the simple test for an asymmetric relation. Both properties require that the machines involved in the test are also able to conduct tests of their own, asking questions, engaging in discussions and assessing the responses of their interlocutors. To test for symmetry, interrogators must be subject to interrogation, breaking down the interrogator-candidate distinction. As I have mentioned, this is already implicit in the experiment since we use people who have already been deemed mentally competent to carry out the test (rather than CAPTCHAS).¹⁵

There are several ways we might choose to run this kind of test. One proposal would be the following. The test would have a round-robin format of conversations, conducted privately and on an individual basis to allow for maximum chicanery and gossip (as Turing encouraged). Ideally, each candidate would be able to speak to the other candidates more than once, making references to prior conversations and what they have learned but there would be no restriction on what participants could discuss. Judgements would be withheld until the whole session is complete when each contestant’s assessments would be gathered and compared. In effect, the experiment would be more like speed-dating than a standard Turing test. After two five-minute rounds of conversation, the contestants, both human and machine, provide their judgments of whether they were speaking to a machine or a person (for n participants, this would take $5n(n - 1)$ minutes). These judgements would then be tallied

¹³ Harnad writes that ‘Turing is dead wrong here’ (Harnad 2008, p. 48) while Bringsjord calls this an ‘absolutely dreadful rejoinder’ (Bringsjord 2009, p. 90).

¹⁴ If $\forall x, y [xRy \rightarrow yRx]$ and $\forall x, y, z [xRy \ \& \ yRz \rightarrow xRz]$ then $\forall x \ xRx$. If R is symmetric and xRy then yRx . And if R is also transitive, then $xRy \ \& \ yRx \rightarrow xRx$. It is important to emphasise that this relies upon the set considering being closed under the relation (and so reflexivity isn’t a redundant property of equivalence relations). The nature of this what we mean by closure is discussed below.

¹⁵ It does conflict with Loebner’s desire to do without human confederates (Loebner 2009).

and a graph constructed from which it would be possible to determine recognition relations.

As with the standard Turing test, the bar here is very high. In plausible human conversations, participants discuss their thoughts and feelings; new ideas arise and are played with and shared imaginative spaces are created. The topic of conversation can shift in unexpected ways. It is this unbounded potential that makes conversation a useful testing ground for artificial intelligence. It requires learning from experience and adjusting one's ends accordingly. Turing expected that 'they would be able to converse with each other to sharpen their wits' (Turing 1950b) and in the end, their output will be worth the same attention as a human mind. Conversation provides us with some limited opportunity to see this process in action.

As some of the properties discussed above take the form of conditionals, we can only test for them by looking at falsifying cases. For example, symmetry is failed when a candidate fails to recognise another agent which recognises them as a thinker (or vice versa). Transitivity is failed when a candidate disagrees with a candidate who they recognise about whether some third candidate is intelligent. Reflexivity is failed when either of the other two conditions are not met. We can use these conditions to distinguish several Turing tests which might identify different levels of intelligence. Some candidates might be judged to be human beings but be terrible at judging others. Others might be good judges though obviously machines. It is possible but unlikely that our experiments will result in neat equivalence classes but instead in more complex structures arising from the recognition relation. These may be of independent interest and it may be possible to construct a notion of 'degrees of intelligence' out of this. We might ask contestants to offer their degrees of belief (credence) that another contestant is a human or to rank other contestants according to this. The advantage of doing this would be that we could state a clear threshold of belief for what counts as being recognised as intelligent by other contestants.¹⁶

It's clear that the question of what counts as success here is more complicated than in standard forms of the test and the test can end with ultimately inconclusive results. I don't think that this is a problem with the test itself but with the complexity of the issue we are considering. At an individual level, an agent 'passes' the test if their beliefs cohere with others but the relations that actually matter are more complicated. The relations we are interested in require more than individual-level analysis.

This test keeps in place Turing's emphasis on a dialogical test of intelligence, his emphasis on interrogators' judgements as evidence and the test's broad anti-reductionism. It also has considerable advantages over the traditional Turing test. First, it integrates multiple strands of research; from natural language processing to sentiment analysis to mind-reading, while still being easier to pass than alternatives like

¹⁶ This is one reason, perhaps, to move the focus from 'humanity' a property with biological connotations to 'intelligence'. We are still left with an asymmetry between humans and machines in this test. If a human fails the test (is not regarded as a human by other contestants), they still get to go home at the end of the day while a machine is not so lucky. The serious question is, could a genuinely thinking, intelligent machine fail this test.

the Total Turing test (or the Truly Total Turing Test).¹⁷ It can't be passed by simply appending any currently existing CAPTCHA to a chatbot.

A second virtue of the experiment is its clear connection to the Social Intelligence Hypothesis; the claim that distinctively human intelligence evolved in part as a result of the growth in the complexity of human societies. While the exact details of the hypothesis vary between models, a core theme is that intelligence evolved in part to assist agents in the practices of signalling group affiliation, determining the honesty of other agents, and both deceiving and identifying deception. Advances in one of these areas of research create an increased selection pressure in others while correspondingly 'an increase in social intelligence selects for further increases by increasing the complexity of the social environment' (Sterelny 2007, p. 720). As the transitivity and symmetry conditions rely upon agents forming a consensus, reciprocal cooperation is incentivised.

As machines get better at passing the test, they must get better at filtering out frauds, thereby creating a mutually enforcing selection-pressure for more intelligent agents.¹⁸ Turing himself explored the connection between research into artificial intelligence and the evolution of natural intelligence, suggesting that the judgment of the experimenter should be identified with the process of natural selection (Turing 1950a, p. 460).

It also addresses another common fear about the Turing test, that there is no activity on the machine's side. One of the reasons a Chinese room appears to be unintelligent is that it merely responds to verbal inputs. The ability to produce an appropriate response to a verbal input is little different to producing an appropriate numerical response to a numerical input. Leaving aside whether or not the room has the appropriate phenomenology of thinking ('a sort of buzzing that went on inside my head' in Turing's words), we're surely right to be suspicious of a system which merely responds to inputs with no conversational objectives of its own. Asking questions has traditionally been a stalling tactic for conversational agents (the most famous example of this is ELIZA, Joseph Weizenbaum's early pattern-matching program which would ask interlocutors questions like 'who laughed at you recently?' and 'can you elaborate on that?') (Weizenbaum 1976). In the revised test, agents ask questions for a purpose; so that they can form a judgment about who they are speaking to, their intentions, and ultimately whether they are intelligent. This is a form of purposive activity. To compete successfully in this, the machine will have to be as reliable as human contestants at identifying intelligence.

As it has not yet been conducted, we must confine ourselves here to assessing its merits as a thought experiment keeping in mind that the success of any given thought experiment is down in large part to the imagination of the reader. I am fairly convinced that if a machine passed this test regularly, we would regard it as intelligent in an important sense. I don't mean to claim that it would have 'full-blown

¹⁷ While it may help if a machine memorised a handful of Winograd schemes or anaphora resolution tasks to pose to its interlocutors, this would only go so far.

¹⁸ One might expect better results from a machine trained with some kind of Generative Adversarial Network, a method of learning which has already lead to advances in CAPTCHA development.

human intelligence' whatever that might mean but I do think that it should be accepted into our community of intelligent agents.

4 Objections to the Reciprocal Test

We'll now consider some objections to the revised Turing test. The two which I find most compelling are the *argument from modularity* and the *argument from logocentrism*. While the first applies specifically to this form of the Turing test (and the arguments I have used to motivate it), the second argument has broader significance. As I believe the appropriate response to these arguments touches on the same themes, I will present the arguments together first before responding to them.

4.1 The Argument from Modularity

One plausible objection to this idea goes something like this. This kind of fuzzy Hegelian nonsense is exactly what proponents of modularity warned us about! (Fodor 1985). Just because a machine has the capacity to reliably judge intelligence and be judged as intelligent does not entail that it has the ability to convince itself that it's intelligent. That symmetry and transitivity entail reflexivity is a neat fact of logic but abstracts away from important features of cognitive architecture. It is perfectly possible that the capacities that enable us to judge intelligence and pass as intelligent are modularised. What would be required for this proof to be relevant is the further claim that the machine is capable of forming some internal representation of itself and recognising that to be intelligent. Without this further claim (which is closer to what proponents of reflexive thought are demanding), the appeals to logic are empty. In any case, even if you had some evidence that the machine is capable of self-training, for example, by running dialogical contexts on itself to enhance its skills, that would not demonstrate that the machine was *aware* that it was itself that it was running these simulations on. In short, the proof of the reflexivity of recognition is unconvincing and so such a test would be unable to tell us if a machine is truly self-conscious.

Further, even if we did accept this form of argument, the claim that a relation is reflexive over a set iff it is transitive and symmetric over that set only holds if we have a pre-existing set over which we are defining relations.¹⁹ The definition of equivalence classes requires transitivity, symmetry and reflexivity; the synthesis of social substance isn't so simple. It's possible that recognition relations don't define neat equivalence classes. In the past, certain groups were not regarded as intelligent humans by people who were regarded as intelligent themselves but we obviously shouldn't endorse these claims. Social constructivist accounts of intelligence are

¹⁹ The relation may define a proper subset of our set. Consider the set $X = \{1, 2, 3\}$ and the relation $R = \{\langle 1, 1 \rangle, \langle 1, 2 \rangle, \langle 2, 1 \rangle, \langle 2, 2 \rangle\}$. This relation is transitive and symmetric but it is not reflexive on the set X as no element relates to 3.

inherently dangerous. It is far better to endorse a psychologicistic theory which lays down clear criteria for what makes something intelligent.

4.2 The Argument from Semantic Externalism

The second objection concerns the familiar charge of logocentrism. We are presenting a linguistic account of recognition. There are two serious objections to this. The first is that language is not necessary for intelligence. I've already mentioned that the Turing test overlooks many skills pertaining to physical embodiment such as spatial reasoning, planning, problem-solving and so forth. It's pretty widely held that physical embodiment may be necessary for intelligence. We will have to stipulate that we are working with a different kind of intelligence here and affirm that intelligence is not monolithic. The second is that physical embodiment is a necessary condition for linguistic reference. Putnam, Davidson and Scweizer all argue that no purely linguistic form of the Turing test would be adequate as embodiment is required for an agent to hook up to the kinds of causal chains which ground linguistic reference.²⁰ If this is the case, then regardless of how an agent performs in the test, its words would not be referring to anything (and perhaps its mental states would lack content).

5 Clarification and Defense

I think these are both strong objections to certain forms of the Turing test but they don't hit their mark here. To see this, it's important to be clear about the analysis that is being proposed. It might look like the idea in this paper conflates two approaches to the issue of intelligence. One approach identifies intelligence with a set of skills, in this case, the ability to form social groups and alliances and the capacities for mind-reading, prediction, and natural language processing that support this. Another approach regards intelligence as socially constructed in the sense that to be intelligent is simply to be accepted as intelligent by members of a community. The first approach merely identifies intelligence with social intelligence while the second proposes a form of social constructivism about intelligence.

These are not exclusive options. I take both of these claims to be true *in a sense* but that they must be allowed to inform each other. To assume that either interpretation is *the* correct interpretation would be to make the kind of mistake that response-dependence theorists caution us against. Recall that response-dependence theorists want us to stop us trying to identify the meaning of a concept with either the stimuli that cause us to token it or the phenomenal properties involved in that tokening.

²⁰ 'What is needed is evidence that the object uses its words to refer to things in the world, that its predicates are true of things in the world, that it knows the truth conditions of its sentences. Evidence for this can come only from further knowledge of the nature of the object, knowledge of how some of its verbal responses are keyed to events in and aspects of the world, events and aspects also known to the interrogator' (Davidson 1990, p. 83). 'There is *some* causal connection between the machine and the real-world apples, etc., via the perceptual experience and knowledge of the creator—designers. But such a weak connection can hardly suffice for reference' (Putnam, 1981, p. 11).

What is being proposed here is that we shouldn't identify intelligence with the internal abilities involved in using the concept of intelligence or with the behaviours manifested that give rise to these judgments. To be more specific, identifying the property of possessing intelligence with the capacity to use the concept of intelligence would be the same mistake as identifying the property of redness with the responses red things produce. Intelligence is not simply the ability to use the concept of intelligence (i.e. the underlying capacities exercised in applying the concept).

What is being claimed instead is that the concept of intelligence is unintelligible except with reference to acts of ascribing and tracking intelligence and that when we ascribe intelligence, this is at least a part of what we concerned with. In much the same way that, to understand concepts like 'red', we had to consider the relationship between an agent and their environment, to understand intelligence, we must understand how the agent responds to their social environment. While red things have the disposition to cause red sensations, the disposition-to-cause-a-sensation-of-redness relation doesn't need to be symmetric. Something can be non-red and yet still experience redness. In the case of intelligence, matters are more complex. To be intelligent is to be the kind of thing that recognises others and is recognised as intelligent. Your intelligence depends upon other minds, in part, because your intelligence constitutively involves the capacity to classify others as having minds.

Similarly, to know that something is red is to know that it is liable to cause a certain response (though to reiterate redness is not identical with that response, it should not be identified with an internal state of an agent). If either agents' responses or the physical properties which give rise to them were different, the concept of red would be different.²¹ In other words, an agent's concept of red depends upon its environment. The same holds for our concept of intelligence. Whether or not an agent has the concept of intelligence depends not only on their internal mental states but their relation to other intelligent agents. And their ability to have this concept depends upon others.

With this in mind, let's return to the two objections above. We'll start with the argument from semantic externalism. The charge here is that our machine wouldn't be appropriately causally connected to the world for its words to mean anything. Specifically, I want to focus on the idea that a mechanical agent would not have acquired its words in a manner that enables them to latch on to the causal chains or social facts which determine semantic reference. As Schweizer puts it, '[I]f 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' (Schweizer 2012, p. 203).

It is important to distinguish two issues here. First, whether an agent must have acquired their language in a manner similar enough to how a human acquires their language in order for us to interpret them as we do a human. Second, whether the

²¹ In a world where our sensation of red was only caused by things which in our world are blue, we would regard the concept of red to be different.

Turing test can tell us this?²² If, when talking about artificial intelligence, we mean that the agent has had its knowledge ‘programmed into it’, then this is a serious problem (assuming some form of socially-oriented semantic externalism is true). But this is not the account of intelligence proposed by Turing. Turing repeatedly proposes that a machine should have undergone an education process similar to humans: ‘The potentialities of human intelligence can only be realised if suitable education is provided. The investigation mainly centers round an analogous teaching process applied to machines’ (Turing 1948, pp. 431–432). A good deal of his writing on machine intelligence concerns the importance of study and enculturation for machines (Turing 1950b). In particular, he proposes that the education of the machine be ‘entrusted to some highly competent schoolmaster’ who would be responsible for teaching the machine English.²³

Turing was explicit that the machine’s knowledge would be caused by experience rather than programmed in by the machine’s creators. If we stick to what Turing proposed then, these concerns about causal chains and the social dimension of meaning miss their mark. It is reasonable to assume that Turing expected any machines entering the Turing test to have undergone this form of education and this would be a sensible demand to make on any machine entering any other Turing test. It is an incredibly difficult demand to satisfy but it provides a clear focus for research. While a machine that has been educated like a human child does seem to be what Turing was talking about, he would presumably still consider a non-traditionally educated machine as intelligent if it passed the test.

It is true that the test doesn’t tell us how an agent acquired its knowledge and so can never truly settle doubts about whether the participant’s words partake in the same causal chains as our own (we have the same problem when we write to pen pals). When I speak to you, I can never know for sure that you’re not systematically confused, from another dimension, or didn’t learn your language from a book that was created by the kind of fortuitous circumstances that gave rise to Swampman.

If judging that something is intelligent depends upon judging what kind of thing it is, then our concept of intelligence will always be species-specific and so not the kind of thing that humans and machines can share. But if we accept that intelligence can be possessed by both humans and machines and that it is *the same thing* in both these instances, then it is reasonable to assume that we can test for it in the same

²² Schweizer argues that an artificial agent would not count as an intelligence if it could only live with a parasitic form of intentionality. ‘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’ (Schweizer 2012, p. 211). One of the purposes of the Reciprocal Test is to show how such sociality could be established.

²³ ‘As I have mentioned, the isolated man does not develop any intellectual power. It is necessary for him to be immersed in an environment of other men, whose techniques he absorbs during the first 20 years of his life’ Turing 1948, p. 432) ‘It must be given some tuition. We need not be too concerned about the legs, eyes, etc. The example of Miss Helen Keller shows that education can take place provided that communication in both directions between teacher and pupil can take place by some means or other’ (Turing 1950a, p. 461). If Helen Keller was able to connect herself to reference-fixing causal chains without being able to see or hear her interlocutors, these sense modalities presumably aren’t necessary for communication.

way. The fundamental idea of the imitation game is that if we understand intelligence to be human intelligence, then it should be tested for in exactly the way that we test for human intelligence (hence Turing's regular appeal to the *viva voce* as a method for establishing whether someone has successfully undergone a process of education or whether they are merely 'parroting' answers). No single test can guarantee that a human is not secretly a machine, a machine is not a Mechanical Turk, or some other sort of mischief is not afoot. This is something we accept in all other forms of examination and there is little reason to think that an intelligence test like the imitation game would be any different.

What if we were to weaken the claim behind the objection? One needn't be a semantic externalist to defend the requirement that an intelligent agent is embodied in some way. One might reasonably argue that a grasp of a language requires mastery of some kind of language-entry and language-exit rules and any machine that lacked sensory facilities would be lacking these. Furthermore, the Turing test, by focusing on lab-based dialogue rather than worldly-engagement, fails to test for these abilities and is, therefore, an inadequate method for determining if a machine is thinking.

I am not wholly convinced that the connection between language and the world needs to be that direct. Consider a neural network that has been trained to identify cats in images. In a sense, it grasps the language entry rules for 'cat'. It doesn't know much else but it can categorise cats and non-cats appropriately. It isn't clear that the machine needs to be able to engage with the cat (e.g. pick it up, pet it...) in order to possess the relevant rules. It does not need to do this in real-time interactions with its environment.²⁴ If we grant that a machine that can identify cats from images possesses the relevant language-entry and exit rules, then it would seem that 'embeddedness' is not required (though it may no doubt be vital for other aspects of cognition).

The first challenge is more explicitly directed at the Reciprocal Test. Even if a machine is repeatedly identified as intelligent while also reliably identifying other intelligent agents, it seems we cannot be guaranteed that it is itself intelligent because that would require a separate ability to identify itself as intelligent and at least some internal representation would be required to support this.

I think there are two decent responses to this charge. I agree with critics that Turing was much too quick to make the solipsism charge. One can demand that self-awareness is a necessary condition for mindedness while denying that we can only have knowledge of others' minds by possessing *their* self-awareness. And while assessing whether an agent is intelligent based on its internal self-perception is obviously not something we do; it does nonetheless seem to be a part of our judgment that an agent is intelligent. At best, the argument given above suggests that it is entailed by the first two judgments. The problem is that we might wish to withhold or retract our initial judgments upon finding out that we have been engaging with a

²⁴ If the lag between light hitting our retina and a 3D image being created were any greater (e.g. 0.5 ms, 1 s, 20 min...), I'm not certain this would radically alter the content of many of the concepts we operate with.

machine. That is, you might find yourself committed to the claim that an agent could identify itself as intelligent but change your mind once the test is over. Such retraction is the norm in Turing tests.²⁵

This suggests that we should not understand the entailment above as a matter of logical deduction but inference to the best explanation.²⁶ Here's what I mean. Like Hume, I find that when I look 'inside', I don't come across any internal representation of myself. I nonetheless distinguish myself from both unintelligent matter and other intelligent beings. What identity I have, I preserve this way and these skills are exactly what would be required of any agent that could pass the Reciprocal Test. Such an agent would need to be able to track its own claims as well as others. It would need to know that its past claims are ones that it can retract and that its future claims will be rationally constrained by what it has already said (what it says must be compatible with both what it has said and what is entailed by what it has said if it is to appear intelligent). In short, it would have to be able to track its past activity and acknowledge that its own past activity differs from both other intelligent agents and unintelligent mechanical competitors. These abilities would be needed to carry out the tasks required to instantiate symmetric and transitive relations and I think we would be entitled to call the sum of these processes a kind of internal self-representation.

The second response brings us back to the issue of the Gallup Poll. If we grant that an agent who passes the Reciprocal Test at the very least has exhibited a grasp of the concept of intelligence and is able to use the *word* as accurately as the rest of us, then they may be as responsible as anyone else for determining its extension. And if they do apply it to themselves, we may not be in a good position to deny them. While Turing dismissed the use of Gallup polls to determine the answer to the question 'can machines think' as absurd, he did nonetheless present his own views as a prediction about future usage. If machines can successfully integrate themselves into our linguistic communities, they will eventually have some role in determining how terms like 'think' and 'mind' are used and if they choose to apply these terms to themselves, we might expect them to be as incorrigible as the rest of us. It may become appropriate to say that machines 'think' because machines play an important role in determining how the word 'think' is used on a daily basis. I don't think we are too far from having the first, widespread, machine-coined neologism and whether this term is applied correctly will be, in part, down to how it is used by a machine.

²⁵ It's worth noting that a person might still respond this way even if the machine is genuinely intelligent. I am not aware of any behavioural test by which I could demonstrate that I am self-conscious. I am taking self-awareness here as being awareness of something like a traditional 'self', some inner subject of experience which endures across time and which may be the referent of the singular first-person pronoun.

²⁶ This is consonant with Turing's discussion of the test as statistical in nature. The idea that the test provides us with an inference to best explanation is not novel (see Shieber 2007 for discussion).

6 Conclusion

The test described in this paper is motivated by several ideas. The first was that the function of the concept of intelligence is to distinguish intelligent from non-intelligent beings ('mind from matter', 'thought from non-thought'). The only necessary requirements of a judge in the Turing test is that they fall under this concept and that they can wield it themselves. They do not need to be experts in machine learning or robotics and while we don't assume that judges have an explicit theory of intelligence, we do assume that they can identify agents like themselves. This provides us with a method of triangulation and it captures why the issue of artificial intelligence is interesting—we want to encounter agents like ourselves. In doing this, we are assuming that if an agent has been identified as intelligent, it too has a grasp of the concept of intelligence—the ability to distinguish intelligences from non-intelligences. This coheres with the social-intelligence hypothesis as it connects intelligence with the ability to track intelligence. It also introduces a selection-pressure on research in artificial intelligence and combines several branches of research into a single, independently-motivated project.

The Reciprocal Test is not perfect and we are far from having machines that could pass it but it captures something important about the concept of intelligence that is left out by other tests. It shows how, if we are externalists not just about the concept of intelligence but the property as well, we can still study how this property emerges in communities. What our term 'intelligence' denotes is something which is, in part, grounded in relations of reciprocal recognition and thus not something which can be understood exclusively in terms of internal cognitive processes.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Block, N. (1981). Psychologism and behaviorism. *Philosophical Review*, 90, 5–43.
- Brandom, R. (2019). *A Spirit of Trust: A Reading of Hegel's Phenomenology of Spirit*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Bringsjord S. (2009) If I Were Judge. In: Epstein R., Roberts G., Beber G. (eds) *Parsing the Turing Test*. Springer, Dordrecht
- Davidson, D. (1990). Turing's test. In K. A. Said, M. Said, W. H. Newton-Smith, R. Viale, & K. V. Wilkes (Eds.), *Modelling the mind* (pp. 1–11). Oxford, UK: Clarendon Press.
- Fodor, J. (1985). Precis of The Modularity of Mind. *Behavioral and Brain Sciences*, 8(1), 1–42.
- French, R. (2000). The Turing test: The first 50 years. *Trends in Cognitive Sciences*, 4, 115–122.
- Haukioja, J. (2007). How (not) to specify normal conditions for response-dependent concepts. *Australasian Journal of Philosophy*, 85(2), 325–331.

- Hayes, P., & Ford, K. (1995). Turing test considered harmful. *Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, 1*, 972–977.
- Pettit, P. (1991). Realism and response-dependence. *Mind*, 100(400), 586–626.
- Loebner, H. (2009). How to hold a Turing test contest. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing test*. Dordrecht: Springer.
- Proudfoot, D. (2013). Rethinking Turing's test. *Journal of Philosophy*, 110(7), 391–411.
- Proudfoot, D. (2020). Rethinking Turing's test and the philosophical implications. *Minds & Machines*. <https://doi.org/10.1007/s11023-020-09534-7>.
- Putnam, H. (1981). Brains in a vat. In *Reason, truth and history* (pp. 1–21). Cambridge: Cambridge University Press.
- Regier, T., & Kay, P. (2009). Language, thought and color: Whorf was half right. *Trends in Cognitive Science*, 13(10), 439–446.
- Schweizer, P. (1998). The truly total Turing test. *Minds and Machines*, 8, 263–272.
- Schweizer, P. (2012). The externalist foundations of a Truly Total Turing Test. *Minds and Machines*, 22, 191–212.
- Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3, 417–424.
- Shannon, C., & McCarthy, J. (Eds.). (1956). *Automata studies*. Princeton: Princeton University Press.
- Shieber, S. (2007). The Turing test as Interactive Proof. *Nous*, 41, 33–60.
- Sterelny, K. (2007). Social intelligence, human intelligence and niche construction. *Phil. Trans. R. Soc. B*, 362, 719–730.
- Turing, A. M. (2004a). (1948) Intelligent machinery, national physical laboratory report, reproduced. In B. Copeland (Ed.), *The essential Turing* (pp. 410–432). New York: Oxford University Press.
- Turing, A. M. (2004b). (1950a) Computing machinery and intelligence. In B. Copeland (Ed.), *The essential Turing* (pp. 433–465). New York: Oxford University Press.
- Turing, A. M. (2004c). (1950b) Intelligent machinery: A heretical theory. In B. Copeland (Ed.), *The essential Turing* (pp. 465–476). New York: Oxford University Press.
- Turing, A. M., Braithwaite, R., Jefferson, G., & Newman, M. (2004). (1952) Can automatic calculating machines be said to think? BBC radio broadcast. In B. Copeland (Ed.), *The essential Turing* (pp. 494–506). New York: Oxford University Press.
- Watt, S. (1996) Naive Psychology and the Inverted Turing Test. *Psychology* 7(14).
- Weizenbaum, J. (1976). *Computer power and human reason*. New York: W.H. Freeman and Company.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.