Large Language Models and the Turing Test:
The “Use of Words” vs. “General Educated Opinion”

William J. Rapaport

Department of Computer Science and Engineering,
Department of Philosophy, Department of Linguistics,
and Center for Cognitive Science
University at Buffalo, The State University of New York,
Buffalo, NY 14260-2500

rapaport@buffalo.edu
http://www.cse.buffalo.edu/~rapaport/

January 8, 2025

Abstract

Passing the Turing Test is often taken to be a sign of intelligence. Some people believe that large language models (LLMs) have passed the Turing Test. Some of those people infer that, therefore, such LLMs show signs of intelligence or thinking. Others of them say that such LLMs, despite passing, are not intelligent (as the Chinese Room Argument holds). And still other people believe that such LLMs have not passed the Turing Test. This essay attempts to sort these issues out, arguing that, even though LLMs may *not* be intelligent, Turing was right: What intelligence “really is” is less important than whether we *take* an LLM to be intelligent. Passing or not passing a Turing Test is irrelevant to the social and moral issues involved with taking the output of LLMs at face value.

"I meant, were these things … well, intelligent? Could they talk?"
"Aye. They could talk. They were intelligent, for-bye,[1] which is not always the same thing."
—C.S. Lewis, *That Hideous Strength* (1946, Ch. 9, §3, pp. 190f)

Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.
—Alan M. Turing (1950, §6, p. 442)

1 Introduction

What will happen when we accept a computer as having passed a Turing test?
—William J. Rapaport (2000, §9, p. 487)

A recent headline in *The New York Times* asked, "Can Apple's iPhones Pass the A.I. Test?" (Mickle, 2024). The "test" referred to wasn't Turing's, but concerned the impact of ChatGPT-enabled software on sales. However, part of the software's success (or failure) will depend on whether users will utilize its output and accept it as trustworthy, which is part of passing a Turing Test.[2]

Some people believe that large language models (LLMs) such as ChatGPT[3] have passed the Turing Test. Some of those people infer that, therefore, such LLMs show signs of intelligence or thinking. Others of them say that such LLMs, despite passing, are not intelligent (as the Chinese Room Argument holds). And still others believe that such LLMs have not passed the Turing Test. This essay attempts to sort these issues out.

I argue that (some) "non-expert" users are "credulous", holding that LLMs have passed Turing Tests and treating them as (if) intelligent, but that (some) "expert" users are "incredulous", holding that LLMs have not passed Turing Tests and (therefore) lack important features of intelligence. Ultimately, however, passing or not passing is irrelevant to the social and moral issues involved with taking the output[4] of LLMs at face value.

[1] Scottish for "besides".

[2] I capitalize 'Test' in this context, because it is not clear that the Turing Test is a "test". Turing, after all, called it a 'game', not a test. Cf. similar remarks about the Turing *Machine*, which is not a (physical) "machine" (Rapaport, 2023, §1.4.1, p. 44, "Terminological Digression").

[3] https://openai.com/chatgpt

[4] I prefer to discuss the "output" of LLMs, because I want to avoid the somewhat theory-laden term 'behavior'.

1.1 The Puzzle and the Problem

Consider the following inconsistent triad:[5]

1. For any *x*, if *x* passes the Turing Test, then *x* can think.

2. Current LLMs such as ChatGPT pass the Turing Test.

3. Current LLMs are "stochastic parrots" that do not think.

Claim 1 expresses a relationship between the epistemological question of how one might *decide* if a computer is thinking and the ontological question of what thinking *is* (Proudfoot 2013; Gonçalves 2021, p. 61). Strictly speaking, the Turing Test is a test of language *use*, which is a sign, or a symptom, or evidence of language *understanding*, which, in turn, is a sign, a symptom, or evidence of intelligence or thinking. Traditionally, it is held that if something passes the Turing Test, then it can think, or is intelligent, or can use language, or understands natural language. A slightly weaker claim is that if something passes the Turing Test, then the "interrogator"—whom I will call the 'judge'—will *take it to be* intelligent (or to think, or to understand natural language). This raises a number of questions: What is "the" Turing Test? What counts as "passing" it? How are language use, (language) understanding, and intelligence related? What is "intelligence" or "thinking"? Is there a difference between being *taken to be* intelligent (or whatever) and *being* intelligent (or whatever)?

Claim 2: The astounding accomplishments—as well as the astounding failures—of multimodal, pre-trained, transformer-based, generative LLMs such as the various incarnations of GPT,[6] ChatGPT, Bard,[7] LaMDA,[8] etc.,[9] have been greeted by the media, pundits, and serious researchers alike as evidence that AI is well on its way to success. Some say that LLMs have passed the Turing Test (Rothman, 2023), or come close to passing, or come "closer ... than any other system to date (although 'closer' does not mean 'close')" (Chalmers, 2020): A time-traveler from the distant past conversing with an LLM would almost certainly take it as intelligent. LLMs seem clearly to be fluent in language *use*,[10] and they "allow for the

[5]On inconsistent triads in philosophical argumentation, see Rescher 1978; Rapaport 1984.

[6]https://openai.com/research/language-unsupervised

[7]https://bard.google.com/chat

[8]https://blog.google/technology/ai/lamda/

[9]Or should that be 'et al.'? Some argue that the 'al.' should only refer to people. 'Alia', however, is of neuter gender. See https://tinyurl.com/websteretal

[10]More specifically, fluent in English (Savage, 2024b). Some are also "fluent" in imagery: "Now, not even reality is required for photographs to look authentic—just artificial intelligence responding to a prompt. *Even experts sometimes struggle to tell if one is real or not. Can you?*" (Hsu and Myers, 2023, my italics). I will focus primarily on language.

creation of content that is often indistinguishable from content created by humans ...” (Bommasani et al., 2022, p. 127). Is that a sign of understanding language? Have they passed the Test? If so, does it mean that they are intelligent, or close to intelligent, or closer to intelligence than any other system to date? What does it mean to be “close” to intelligence? Is complete indistinguishability from a known intelligence necessary?[11] What if the closeness is the result of intentional or unintentional fooling or trickery?

Claim 3 has two parts:

3a. Current LLMs are “stochastic parrots” (Bender et al., 2021).
3b. Stochastic parrots don’t think.

As for claim 3a, we will look at what LLMs do in §2. As to claim 3b, there is a difference between passing a test and *using* one’s knowledge: After all, even if LLMs *can* pass medical school tests, they do *not* necessarily give good medical advice.[12] Although LLMs are generally acknowledged to be amazing, they are not generally acknowledged to *be intelligent*. Does this mean that the (strong) Turing Test conditional (claim 1) is false or has been falsified (as in John Searle’s Chinese Room Argument)? Or do LLMs *not* pass the Turing Test? A system that can communicate with you in English will certainly *seem* to be intelligent. And some people certainly *take* LLMs to be intelligent. Are they, indeed, intelligent? Do computers really understand what they are doing? Do *we* understand what they are doing? Do we understand what *we* do? Do we understand how our brains produce our intelligence? Does it matter how intelligence is produced?

The indistinguishability of the output of LLMs from “content created by humans” is part of the puzzle: After all, isn’t creating such indistinguishable content what the Turing Test is all about? It is also part of the problem: After all, if LLMs are not intelligent, then it becomes important to be able to distinguish their output from ours. It is also important to do that even if they *are* intelligent:

> It is easy to get seduced by the artificial intelligence of GPT-4. It can ace the bar exam! It can get perfect scores on Advanced Placement tests! It knows how to code! ... But soon, too, it will be able to generate seamless deepfakes and create images from text including, no doubt, pictures of child sexual abuse. It is a powerful, seismic technology that has the capacity both to enhance our lives and diminish them. Without guardrails and oversight, its harms are destined to multiply. (Halpern, 2023)

[11] Note that, although a printed reproduction of a (photo of a) painting is not *completely* indistinguishable from the original painting, it is indistinguishable *enough* to serve as a substitute in many contexts (such as art books).

[12] Singhal et al. 2023; Tamayo-Sarver 2023; Marcus 2024i. But see Goh et al. 2024; Kolata 2024 for contrasting evidence.

1.2 Background Assumptions and a Look Ahead

Despite the popular press's current use of 'AI' to refer to certain computer programs or to current deep-learning methodologies, AI is more accurately characterized as the branch of computer science that investigates whether (and the extent to which) cognition is computable (Barr 1983; Rapaport 1998, 2020; Rapaport 2023, Ch. 18). For present purposes, I will assume that there is no logical or *a priori*—as opposed to practical—reason to think that AI cannot accomplish its goal of developing a computational theory of cognition: No one has logically proved that cognition is not computable in the way that the Halting Problem is not computable (Rapaport, 2024b). (One caveat: As Keith Gunderson (1985, p. 180) notes, 'Can machines think?' is not a yes-no question. Neither is 'Is cognition computable?', because 'cognition' is vague.)

To what extent are LLMs a step in the direction of computational cognition? Must a computable theory of cognition be in any way similar or analogous to the inner workings of *human* cognition? At one point, Herbert Simon seemed to think so: "the *processes* some of these [GOFAI][13] programs use parallel closely the observed human processes" (Simon, 1996, p. 7, my italics). It is not at all clear that LLM processes "parallel closely" the human ones.[14] The processes that Simon had in mind were very general ones "that can create and operate upon" "patternings of matter" (p. 7). That could include both computational and biological processes. Elsewhere, he allowed that the processes could either be "humanoid ... or by *brute force*" (quoted in Hearst and Hirsh 2000, p. 8, my italics). The question is what "patterns" (or data) and what "processes" will do the trick. But human cognition might be computable even if not computational (Rapaport, 1998). The successes of LLMs (such as they are) suggest that many features of intelligence are embedded in our recorded linguistic texts (the LLMs' input) and need not be due to cognitive processing. Rather, they can emerge[15] from what Turing (1952, p. 500) called "donkey work" and what Bender et al. (2021) called "stochastic par-

[13]"Good Old-Fashioned AI" using logical, symbolic, and knowledge-based techniques, as opposed to neural-network systems (Haugeland, 1985).

[14]Though see Wolfram 2023 for some observations to the contrary, e.g.: "[H]ow is it ... that ... ChatGPT can get as far as it does with language? The basic answer ... is that language is at a fundamental level somehow simpler than it seems. And this means that ChatGPT—even with its ultimately straightforward neural net structure—is successfully able to 'capture the essence' of human language and the thinking behind it. ... ChatGPT has somehow 'implicitly discovered' whatever regularities in language (and thinking) make this possible" (p. 59). Cf. Buckner 2024 for a book-length discussion of the cognitive capacites of deep-learning systems, including LLMs.

[15]On the existence of "emergent abilities" in LLMs, see Bommasani et al. 2022, pp. 1, 3; Wei et al. 2022. On the *non*-existence of such emergence, see Carter 2023; Schaeffer et al. 2023. On emergence and the notion of unintended consequences, see Hutson 2023.

roting".[16] LLM processes seem to lie towards the brute-force end of the spectrum. Does the processing matter? I will argue that it may not (§3.6).

I will also assume that the Turing Test—better: Turing-*like* tests (plural)—is a good measure of AI accomplishments (Rapaport, 2000, 2024b). But it is also important to keep in mind that the Turing Test is a *replacement* for questions about thinking and intelligence. The replacement is justified to the extent that the computer's output is indistinguishable from human output. Thus, what counts is the system's output, not its internal processing (except to the extent that the internal processing must produce indistinguishable output). However, humans generally take linguistic output at face value, and *assume* that whatever produced it can think or is intelligent. We *attribute* intelligence to the system on the basis of its output. A Turing Test of an AI system is a good measure of its accomplishments to the extent that we are willing to treat the system's output as we would a human's.[17] As a result, some *non-experts* (ordinary people) have judged that LLMs *have* passed the Turing Test and therefore *treat* them *as* intelligent. And therein lies a serious problem, not least because of LLMs' inability to distinguish fact from fiction—not to mention a *judge's* inability to make that distinction! We'll consider these points in detail in §3.

But there are aspects of the output that cause some *experts* in AI (or cognitive science, or intelligence) to judge that LLMs have *not* passed a Turing Test: LLMs don't (yet) exhibit important aspects of cognition. We'll discuss these in §6.

To think about this more deeply, we need to look at what LLMs do (claim 3a) and revisit the Turing Test (claim 1). That should help us think about whether LLMs pass (claim 2) and whether they think (point 3b). Let's begin with LLMs.

2 The Nature of LLMs

> What does a bowl of alphabet soup know?
> —"Ziggy" cartoon (17 September 2002), https://tinyurl.com/ziggy20020917

2.1 Terminology

As I understand the recent literature, the terminology for these language-processing systems is not yet fixed. Following Bender and Koller (2020, p. 5185, col. 1), let's

[16]Note, by the way, the role of non-human animal metaphors (shades of Descartes!).

[17]Ferrario et al. 2024 argue that AI systems by themselves are neither "experts" nor "authorities", and that only hybrid systems consisting of humans working together with AIs can have such "epistemic superiority". Goh et al. 2024 provides evidence to the contrary.

say that a *language model* (LM)[18] is any neural-network machine-learning "system [based on "transformer" technology[19]] trained only on the task of string prediction, whether it operates over characters, words or sentences . . . ". Then a *large* language model (LLM) is one based on (very) large amounts of training data.[20] Another term sometimes used to generalize over these (so as to include non-linguistic systems such as DALL-E[21]) is 'foundation model' (Bommasani et al., 2022).

Murray Shanahan (2024) distinguishes between (1) LLMs as "generative mathematical models of the statistical distribution of tokens in the vast public corpus of human-generated text" (p. 70, col. 1) and (2) "the systems in which they are embedded" (col. 3), such as GPT, ChatGPT, Bard, etc. This is an important distinction, but I will usually ignore it here (as does Shanahan occasionally; see p. 68). For convenience, I will call them all 'LLMs'.

2.2 What Do LLMs Do?

To a first approximation, an LLM does "super-autocomplete" (Chomsky et al., 2023). It is . . .

> . . . always fundamentally trying to . . . produce a "reasonable continuation" of whatever text it's got so far, where by "reasonable" we mean "what one might expect someone to write after seeing what people have written on billions of webpages, etc." . . . So at any given point, it's got a certain amount of text—and its goal is to come up with an appropriate choice for the next token to add. (Wolfram, 2023, pp. 1, 47)[22]

One of the things that is fascinating about LLMs is the difference between their internal processing and their linguistically fluent and intelligent-seeming output:

> Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot. (Bender et al., 2021, pp. 616–617)

Although this is a nice statement of what LLMs "really" do "at bottom", remember that what *humans* "really" do "at bottom" is fire neurons[23] (and much else besides,

[18]Although LMs may be "models" of language, this use of 'model' must be distinguished from the notion of a model of the *world*, as discussed in §6, point 10, below.

[19]Vaswani et al. 2017; see also Levinstein 2023a.

[20]https://openai.com/research/better-language-models

[21]https://openai.com/index/dall-e-3/

[22]For excellent overviews of how LLMs work, see Gubelmann 2022, §2, and Levinstein 2023a.

[23]See, e.g., Cappelen and Deaver 2021, pp. 12, 16, 48, 71.

but for simplicity I will lump all of the electrochemical behavior in the brain together as neuron firings). So we might say, in the same vein, that contrary to how it may seem when we observe *our* linguistic output, a human brain is a system for firing neurons. But almost everyone would agree that *our* neuron firings *do* yield meaning. Are we wrong? Or could stochastic parroting also yield meaning?

There are two ways to respond to these observations: In both cases, we can choose to focus only on the internal processing—stochastic parroting and neuron firing. Or, in both cases, we can choose to focus on the external output (the "behavior"), i.e., on apparent machine intelligence and real human intelligence.[24]

2.3 LLMs and AI Research

It may help to ask where LLMs fit into AI research. Consider Stuart C. Shapiro's classification of three different kinds of AI research (1992, p. 54):[25] LLMs are an almost classic example of (1) *AI as advanced computer science or engineering*, extending "the frontier of what we know how to program" by whatever means will do the job. They also fit the category of (2) *AI as computational philosophy*, investigating the extent to which cognition is computable. They are certainly a large step in this direction, even if not the only or the best ones.

But they do not seem to fit the category of (3) *AI as computational psychology*, where AI programs are considered as theories or models of human cognitive behavior. There is little attempt to have them attack a cognitive task such as language understanding *in a cognitive fashion*. Although the underlying technology of neural networks may have been inspired by the biological neural network of the brain,[26] it is generally agreed that this is more metaphorical than literal.

2.4 LLMs and Computational Linguistics

Given their emphasis on linguistic abilities, where do LLMs lie on the spectrum of things studied by computational linguistics? We can distinguish three levels:

- At the lowest level, there is *language data* processing, defined as "the processing of natural language data by computer" (Garvin, 1985). However, in conversation, Garvin once emphasized to me that he considered this more generally as the computational processing of linguistic data, not necessarily considered as "language". On Garvin's view (as I understand it), *language data* processing is not necessarily *natural-language* processing, because the

[24] Although Gunderson (1985, pp. 170, 174) suggests that they should be treated differently.

[25] For discussion of these three views, see Rapaport 2023, Ch. 18. See also Searle 1986.

[26] For a good historical survey of neural networks, see Perconti and Plebe 2023.

linguistic data is not necessarily "natural". LLMs clearly do language data processing.

- At an intermediate level, there are the two aspects of *natural language* processing:[27]
  - natural-language *input* processing,[28] i.e., "reading" and "hearing", preliminary to any understanding, and
  - natural-language *output* processing, i.e., natural-language *generation*.

  LLMs plausibly do natural-language processing.

- At a high level, there is natural-language *understanding*, by which I mean "real" understanding of language the way you are understanding these sentences right now (though what, exactly, *that* might mean is exactly the philosophical issue at hand!). Neither language data processing nor natural-language processing is necessarily natural-language *understanding*. It is not at all clear that LLMs *understand* natural language.

2.5 Syntax vs. Semantics

One important aspect of the construction and behavior of LLMs is a focus on what I will call 'syntax' as opposed to 'semantics'.

By 'syntax', I mean more than mere grammar. I mean to include *any* "internal" properties of, and relations *among*, the tokens, words, sentences, etc., that LLMs deal with, in contrast to any "external" relations with the external (or "real") world (including users). On this view, both distributional "semantics"[29] and conceptual-role "semantics" (Rapaport, 2002) are actually syntactic.

And by 'semantics', I mean exactly those external relations *between* the system and the world. This use of 'semantics' does not include any "internal" (including

[27] Or what Shapiro and Rapaport (1991) once called 'natural-language competence'. The phrase 'natural language' is ambiguous. One meaning is languages such as English, French, etc., occurring naturally in ordinary conversation, writing, etc. But the phrase can also stand in contrast to artificial or made-up languages such as computer-programming languages, Esperanto, or Elvish. Note that some spoken Esperanto learned as a first language is arguably "natural" in the first sense (Lindstedt, 2006; Stria, 2015).

[28] There does not seem to be a good term for this. The obvious candidate, 'natural-language understanding', won't work in the present context.

[29] "The distributional hypothesis is that words that occur in similar contexts tend to have similar meanings" (Turney and Pantel, 2010, p. 142)

distributional and conceptual-role) "semantic" relations *among* the system's own "symbols" (the tokens, etc.).[30]

In these terms, LLMs operate syntactically, not semantically.[31] Token prediction is primarily syntactic in this sense:

> Predicting *what word comes next* is not the same as predicting *what forms a true statement*. ... Consider ... whether you are more likely to see the phrase "pigs fly" or "pigs walk." (Michael, 2020, original italics)

It follows that users have no reason to trust in the (perhaps accidental) truth of what such chatbots say. Note that neither natural-language *processing* nor natural-language *understanding* have anything to do with the truth or falsity of the output. Neither statistical likelihood, grammatical acceptability, nor even semantic acceptability entail truth. The output of an LLM is a fragment of Borges's (1941) Library of Babel or Quine's (1987) "universal library".[32] We'll return to this in §3.7.

3 The Nature of the Turing Test

> On the Internet, nobody knows you're a dog.
> —Peter Steiner (1993), *The New Yorker* cartoon, https://tinyurl.com/steiner1993[33]

> Increasingly, we're surrounded by fake people. Sometimes we know it and sometimes we don't.
> —Matthew Hutson (2023); cf. Dennett 2023b

3.1 Machine? Or Thinking Thing?

In C.S. Lewis's novel *Perelandra*, a human visiting Venus encounters a Venusian that

> came right up and began nudging him with its cold snout about his knees. He was in great perplexity. Was it rational and was this how it talked. Was it irrational but friendly—and if so, how should he respond? ... Or was it merely scratching itself against him? (Lewis 1944, Ch. 4, p. 46)

[30] 'Symbol' carries some baggage that I want to bracket: The symbols of an LLM need not *represent* anything (Rapaport 1995, §2.2.1, p. 55; Rapaport 2023, §16.11.2, p. 385). For more on this view of syntax and semantics, see, e.g., Morris 1938; Posner 1992; Rapaport 1986b, 1988, 2017.

[31] Titus 2023 is an important argument that LLMs as currently implemented do not "possess semantic understanding".

[32] On LLMs and the "Infinite Monkey Theorem" (https://tinyurl.com/55h53fte), see Nazaryan 2024.

[33] See also https://tinyurl.com/hafeez-cartoon

If a robot powered by an LLM "came right up and began" a conversation, would you say that it was rational ("intelligent")? Or irrational (unintelligent) but friendly? Or merely stochastically parroting?

Turing (1950) "replaces" "the question, 'Can machines think?' " with this:

> Let us fix our attention on one particular digital computer *C*. Is it true that ... with an appropriate programme, *C* can be made to play satisfactorily the part of A [the man] in the imitation game, the part of B [the woman] being taken by a man [i.e., a human]? (§5, p. 442)[34]

Recall why Turing proposed this machine version of his Imitation Game—what has come to be known as the Turing Test: He felt that defining " 'machine' and 'think' ... to reflect ... the normal use of the words ... is dangerous" (Turing, 1950, §1, p. 433). There are two reasons for this: Turing's explicit reason is that "a statistical survey" of how the words are used "is absurd". An implicit reason may be that, at least up until 1950 or so, 'machines can think' was an oxymoron.[35] Note that to say this is to embrace a form of (Cartesian) dualism: On one hand, there are physical machines; on the other, there are thinking things. Turing avoids this by claiming that a certain kind of machine—a digital computer—might be able "to give a good showing in the [imitation] game" (§3, p. 436).

In order for the Turing Test to replace the question 'Can machines think?', its machine participants must "be *regularly mistaken* for human beings" (Colombo

[34] Where the context is clear in what follows, only section and (original) page references to Turing's essay will be given. It is interesting to observe that, whereas in Turing 1936 'computer' referred only to humans, here he distinguishes between 'digital computers' and 'human computers' (§4, p. 436), characterizing the former using a highly simplified description of a Turing Machine. So, by 1950, the "use of the word" 'computer' had "altered" to mean a genus with at least two species.

[35] At least one of the *Oxford English Dictionary*'s definitions of 'machine' almost explicitly rules out thinking. See its definitions of 'machine' at http://www.oed.com/view/Entry/111850, senses IV.6.b and V.8.b (a sense in which a machine could be a *person* who acts *unthinkingly*), and especially of its cognate 'mechanical' at http://www.oed.com/view/Entry/115544, sense A.II.7 ("acting or performed without thought"). Similarly, the 1958 Merriam-Webster *Second Unabridged Dictionary*'s fifth definition of 'machine' specifically says that machines are "unintelligent" (i.e., by definition) (Neilson et al., 1958, p. 144, col. 3). See also Sieg 2008, p. 527, fn. 1; p. 574; Gonçalves 2024. Wittgenstein (1934, p. 47) said that "the sentence, 'A machine thinks (perceives, wishes)', seems somehow nonsensical. It is as though we had asked 'Has the number 3 a colour?' "; but cf. Proudfoot 2024. For a discussion contemporaneous with Turing 1950, see Mays 1952, especially this passage:

> The O.E.D. definition does bring out one thing at least, a machine is usually thought of as something which does not possess a private life of its own, it does not indulge in reverie when at its task, it lacks consciousness, intelligence and will. (p. 149)

Although 'machines think' may have been an oxymoron ca 1945, people have been trying to get machines to think since at least the 18th century (Lepore, 2024).

and Piccinini, 2023, §2.2, p. 6, my italics). Evidence for this interpretation is in §1 (pp. 433–434): "It is A's [the man's, or the machine's] object in the game to try and cause C [the interrogating judge][36] to make the wrong identification" of A as the woman (or the human). But such cases of intentional mistaken identity may be of less importance than a related but distinct goal: Turing notes that the game "has the distinct advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a [hu]man.[[37]] No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin" (§2, p. 434). That is, Turing *is* claiming that (or at least suggesting that, or asking whether) it is possible to produce computable output that is "indistinguishable" from human "intellectual capacities".

3.2 Thinking? Intelligence? Understanding?

Turing proposed the machine version of his Imitation Game as a replacement for the question whether machines could *think*. Yet he titled his paper "Computing Machinery and *Intelligence*", suggesting that he considered thinking and intelligence to be more or less synonymous (if equally vague). The vast literature on the Test—as well as the occurrence of the word in the phrases 'artificial intelligence' and 'machine intelligence'—takes the Turing Test to be a measure of intelligence, quickly followed by a caveat that 'intelligence' itself is in need of definition. Curiously, however, besides the essay's title, the words 'intelligence' or 'intelligent' appear in only two other places: once referring to the "intelligence" that a *programmer* might need to improve a program's speed (§7, p. 456), and once to contrast "the completely *disciplined* behavior involved in computation" (my italics) with (undisciplined?) "intelligent behavior" (§7, p. 459). And Turing talks about "understanding" in only two places: once when discussing the similarity of the Imitation Game to *viva voce* (i.e., oral) examinations of "understand[ing] something" (§6(4), p. 446), and again when discussing teaching the computer "to understand and speak English" (§7, p. 460). (We'll come back to the *viva voce* version in §4.) His most frequently used terms are 'intellect' and 'intellectual capacities'.

Granted all of this, I will use 'Intelligent', with a capital 'I', to cover all such vague terms as 'think', 'intelligent', 'intellect', etc.[38] In any case, the important point is that the Turing Test is not a "test" *of* Intelligence; it *replaces* questions about Intelligence.

[36] Note that, in Turing 1950, §1, (roman) 'C' represents the *judge*, but, in §5, (italic) '*C*' represents the *computer*.

[37] Is this a dualistic move, dividing body from mind? Or is it an anti-dualistic claim that (some) bodies *can* exhibit mental qualities?

[38] Buckner 2024, §2.4, adds 'rationality' to this list.

3.3 Language and Intelligence

But what exactly is the Turing Test a test of? Natural-language understanding? Or Intelligence? Programs such as LLMs that participate in Turing Tests arguably do no more than linguistic data processing. So the real question is: Can natural-language understanding (whatever that might be) emerge from linguistic data processing?

What is the relationship between natural-language understanding and such things as understanding in a more general sense or Intelligence? Are there aspects of Intelligence that are independent of natural-language understanding? Of course, to count as Intelligent, a system doesn't have to be Intelligent on *all* counts: As Daniel C. Dennett observed,

> The assumption Turing was prepared to make was this: Nothing could possibly pass the Turing test by winning the imitation game without being able to perform indefinitely many other clearly intelligent actions. ... Maybe it wouldn't do everything we hoped—maybe it wouldn't appreciate ballet, or understand quantum physics, or have a good plan for world peace, but we'd all see that it was surely one of the intelligent, thinking entities in the neighborhood.
>
> ... any computer that actually passed the Turing test would be a thinking thing in every theoretically interesting sense. (Dennett, 1985, pp. 124, 140)

That is, a Turing Test-passing computer would have to be "AI-complete" (Shapiro, 1992, pp. 56–57).

This is consistent with one of Stevan Harnad's (1991, p. 44) arguments for a "total" Turing Test:

> The candidate must be able to do, in the real world of objects and people, everything that real people can do, in a way that is indistinguishable (to a person) from the way real people do it.

For this, Harnad believes that the linguistic symbols must be "grounded" in non-linguistic things:

> It is hard to imagine, for example, that a Turing Test candidate could chat with you coherently about the objects in the world till doomsday without ever having encountered any objects directly—on the basis of nothing but "hearsay," so to speak. Some prior direct acquaintance with the world of objects through sensorimotor ... interactions with them would appear to be necessary in order to ground the candidate's words in something other than just more words. (Harnad, 1991, p. 46)

(We'll return to this, and we'll consider Bender and Koller's (2020) octopus example, in §5.3.2.)

Yet, even without such grounding or the ability "to perform ... other clearly intelligent actions", LLMs seem able to pass the Test. To the objection that a thinking thing would have to be able to perceive the external world, or to live in the world, or to have beliefs, desires, and intentions, Dennett says "that the Turing test is so powerful that it will ensure indirectly that these conditions, if they are truly necessary, are met by any successful contestant" (p. 141). Thus, a total Turing Test that actually tests for *doing* things (in addition to merely *talking* about them) is not necessary, because talking will suffice. (See §5.3.2.)

3.4 *Viva Voce* Tests

Note that the kinds of Turing Tests that LLMs participate in are typically of the *viva voce* variety. Turing suggested that the imitation "game **(with the [human] player B omitted)** is frequently used in practice under the name of *viva voce* to discover whether some one really understands something or has 'learnt it parrot fashion' " (§6, p. 446, my boldface). (Think of doctoral dissertation defenses!) Some (e.g., Saygin et al. 2000) have distinguished the single-entity, *viva voce* version of the test from the Imitation Game version, which involves two entities, one of whom is a control (the woman in the Imitation Game, the human in the computer version).

But the *viva voce* version has an implicit control.[39] There are two ways in which an entity (computer or human) can "take a Turing Test"; Dennett (2023a, p. 270) calls them "aggressive probing" and "friendly conversation":

Aggressive Probing: Suppose that the judge is an AI expert, that the test is a sophisticated experiment, and that the judge *can* distinguish the interaction from that of a human (because the entity is lacking in one or more of the features that experts consider to be part of Intelligence). I will call such a judge "incredulous". In that case, an "ideally" Intelligent human is the implicit control, and the entity being judged has not passed.[40]

Friendly Conversation: But suppose that the judge is an ordinary person, that the test is a simple interaction, and that the judge cannot (or merely does not) distinguish the interaction from that of a human. I will call such a judge "credulous". In that case, other ordinary people are an implicit control group, and the entity being judged has passed and has been taken by that judge to be Intelligent. (Note that the judge does not even have to be aware that he or she is conducting a Turing Test.)

[39]For more on the notion of a control in the Turing Test, see Gonçalves 2023.

[40]For observations on how to judge a Turing Test incredulously, see Singhal et al. 2023, p. 178.

Turing himself suggested that passing only required "an average interrogator" (§6, p. 442). Indeed, two years later, he went further and said that "A considerable proportion of a jury, *who should not be expert about machines*, must be taken in by the pretence" (Turing et al. 1952, p. 495, my italics; cf. Proudfoot 2013, p. 398). However, that does not necessarily mean that Turing required his judge to be what I am calling "credulous", because—on my view—I am more concerned with the judge's expertise on *Intelligence*, not necessarily on *computers*. Turing Tests judged credulously are, perhaps, too easily passed! Bisk et al. (2020, p. 8723, col. 2) claim that "Turing was careful to show how easily a naïve tester could be tricked", which suggests that for the Turing Test to be valid, the judge *should* be an expert, despite Turing's preferences.

These distinctions strongly suggest that it is the judge's reactions to the test subject's output that is the determining factor. Whether the output is gobbledygook or fluent and appropriate English, it is the judge who makes the decision (which is not to say that the "decision" is always a conscious, reasoned one!).

3.5 Intelligent Entities

What classes of entity might be candidates for Intelligence? Humans, of course. Other animals, too; apes, corvids, and many others exhibit aspects of Intelligence, even if they cannot pass a linguistic Turing Test (see §5.3.4, below).[41]

Echoes of the debate whether computers can be Intelligent can be heard in the debates whether plants can be. In a review article, Kolbert 2024, p. 6, col. 2, says that "Plants communicate with one another" when one "releases chemicals that prompt its neighbors to beef up their defenses". Is that communication in the sense that people use the word? Or is it a metaphor, as in saying that a certain plant "can, *in a manner of speaking*, hear" (ibid., col. 3, my italics)? Of course, the use of words can alter; such metaphors can become educated opinion! (Rapaport, 2000). Kohler goes on: "What do these structures and strategies reveal about the

[41] Andrews and Huss 2014 distinguishes between anthropomorphism—the (possibly mistaken) attribution "of a human property to a non-human animal" (p. 2)—and "anthropectomy"—the (possibly mistaken) *failure* to attribute a human property to a non-human animal (p. 7). Both notions are highly relevant in the present context if the "human property" in question is Intelligence and if we replace 'non-human animal' with 'AI system'.

Coghlan 2024, p. 3, argues against anthropomorphism, understood as "the claim that people tend to form the . . . false belief" "that some machines really or truly have phenomenally experienced feelings". He cites as evidence that we don't form that belief in the case of puppets or fictional characters. (On dolls—hence also puppets—see O'Connor 2024.) But in those cases, we know (or antecedently believe) that they have no inner mental life, whereas in the Eliza Effect (to be discussed in §3.8.1, below), we don't. Curiously, Coghlan nowhere mentions Dennett's Intentional Stance (see §3.8.3, below). On all of this, see also Astobiza 2024; Buckner 2024, §2.5.

inner lives of plants? Many biologists say not much. ... there is ... no inner life to analyze. Others argue that this is a form of prejudice ...” (ibid.). From one perspective, if the releasing of chemicals is not communication for plants, then statistical processing for computers, and neuron firings for humans, shouldn’t be either. From another, if it is communication for humans, then is there any reason to deny it for plants or computers? And keep in mind that to “take in information and respond to it ... ‘is the basic definition of intelligence’ ” (ibid., col. 4, quoting a plant biologist).

But the Turing Test, and AI in general, is typically and historically aimed at *human*-level Intelligence. When considering the kind of machine that might be tested, Turing specifies digital computers and excludes humans “born in the usual manner”. AI-programmed computers, therefore, are the main non-human candidates.

Turing implies that digital computers can pass a Turing Test for human *computation*:

> The reader must accept it as a fact that digital computers can be constructed, and indeed have been constructed, according to the principles we have described, and that *they can in fact mimic the actions of a human computer very closely.* (§4, p. 438; my italics)

This is as close as Turing comes to acknowledging that his work on the Turing Machine can be considered to be part of AI.[42]

There is one other non-human, non-animal candidate that we should keep in mind: exobiological (i.e., non-robotic) extra-terrestrials. Granted, we don’t know of any, and perhaps there are none. But suppose the fabled extra-terrestrials of science fiction do arrive: How would we determine whether they are Intelligent? Their very arrival would be one sort of evidence, of course. But suppose that, after figuring out how to communicate with them (as in Chiang (2002)), we determine that interactions with them are exactly like our interactions with LLMs. I think we would have to give the same answer to the question of whether the extra-terrestrials are Intelligent as we would to the same question about an LLM. Michael Dummett (1976, p. 70, my italics) observed that

> If a Martian could learn to speak a human language, or a robot be devised to behave in just the ways that are essential to a language-speaker, *an implicit knowledge of the correct theory of meaning for the language could be attributed to the Martian or the robot with as much right as to a human speaker, even though their internal mechanisms were entirely different.*[43]

[42] See Copeland 2023, esp. p. 20, col. 2; Rapaport 2023, §8.7.9, esp. p. 174.

[43] Cf. a similar observation in van Inwagen 1984, pp. 18–19.

LLMs seem to be such "robots". And (some) users *do* seem to attribute such knowledge to them. But it does not follow that they *should*. Although we would probably say that the extra-terrestrials were Intelligent, we would also have to be very cautious in our dealings with them (as we must be with LLMs!).

3.6 Internal Processing? Or External Output?—Part I

Although much recent discussion has focused on LLMs' statistical processing, it is the system's external output that is the candidate for being judged indistinguishable from human output. The judge has no access to the internal processing; all that matters for being judged equivalent is the output (Barr, 1983, p. 5), at what might be called the "psychological" level (as opposed to the neural level in the case of humans and the statistical level in the case of LLMs).[44]

This is precisely what Turing argued: To ask whether the system is "really" Intelligent is beside the point. All that matters is the output and the judge's response to it. All that matters is output equivalence: the indistinguishability from a "control" that is a "gold standard" instance of Intelligence. To my mind, the central passage of Turing's essay is his "use of words and general educated opinion" sentence. If indistinguishability is achieved, then the use of words will alter to recognize that. Whether "general educated opinion" will also alter remains to be seen (we'll consider this in §6).

Are the intellectual abilities of LLMs and humans the same despite their (apparently?) different mechanisms? Turing considers—and rejects—the objection that "machines [may] carry out something which ought to be *described* as thinking but which *is* very different from what a [hu]man does" (§2, p. 435). Different in what way? Clearly not in terms of output; otherwise, the machine would not pass the Test (win the Game). So the differences must be in the "internal" processing—the algorithms that yield the Intelligent output from the input prompts. Turing's response—"if ... a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection" (p. 435)—seems to be that *how* the machine works is irrelevant to *what* it does. In other words, internal processing differences are irrelevant to external output similarities.

How important is the nature of the machine's internal processing? Consistent with the view of AI as engineering, Turing placed no limitations on the nature of the algorithms: "We should wish to permit every kind of engineering technique to be used in our machines" (§3, p. 435). So both symbolic GOFAI and deep-learning techniques should be allowed. (I assume that most LLMs use only very advanced versions of the latter, and none of the former, though GPT-4 may use some GOFAI

[44] We'll look at the analogous distinction between Dennett's (external) Intentional Stance and (internal) Design and Physical Stances in §3.8.3.

techniques.) To the extent that the goal of AI is to develop a computational theory of cognition, it simply does not matter what the internal workings are, as long as they are computable and accomplish the task. On this view, the issues of donkey work and stochastic parroting are red herrings. We'll return to these issues in §4.3; for now, let's focus on the output.

3.7 Fictional and Non-Fictional Language

> It is more from carelessness about truth than from intentional lying, that there is so much falsehood in the world.
> —Samuel Johnson (as quoted in Boswell 1791, Vol. 2, pp. 189–190)

> You could compile ... the worst book in the world entirely out of selecting passages from the best writers in the world.
> —G.K. Chesterton, "An Edition of Dumas" (1907, p. 4, col. 3)

> I dinna believe something only because someone's set words down in a book—for God's sake, I print the damn things! I ken verra well just what charlatans and fools some writers are—I see them! And surely I ken the difference between a romance and a fact set down in cold blood!
> —Diana Gabaldon, *Voyager* (1994, p. 884)

> The Previous [inhabitants] had many other customs that were inexplicable, none more so than their propensity to intermingle fact with fiction, which made it very hard to figure out what had happened and what hadn't.
> —Jasper Fforde, *Shades of Grey* (2009, pp. 7–8)

One aspect of concentrating on the linguistic output that we need to look at is its uniformity or "flatness". Language, *per se*, is syntactic (in the sense of §2.5). As such, it is neither fictional nor non-fictional. By 'fictional language', I mean the language used not only in literary fictional stories, novels, movies, and plays, but also by liars and con artists, in "fake news", etc. And by 'non-fictional language', I mean the language used in non-fiction works, newspapers, science, philosophy, etc.[45] That is, non-fictional sentences don't announce themselves as such; there is no obvious grammatical or syntactic distinction between fictional and non-fictional language (but see below). Arguably, everything that current LLMs say is fiction, even if it's true (Rapaport, 1985).[46]

[45] My original term for this was 'factual language', but 'non-fictional' is a bit more general. It also suggests that fictional language is the default.

[46] "No character in a book is a real person. Not even if he is in a history book and is called Ulysses S. Grant" (Scholes, 1968, p. 17). "All facts presented by Generative AI—even those that are true—are fictitious" (Gerben Wierda, in Marcus 2023a).

Despite this, the default assumption is that language tells the truth, defeated only when there is countervailing evidence, such as language known to be fictional, inconsistency with known facts, or known lying. LLMs' tendency to "hallucinate"[47] is facilitated by the lack of distinction between fictional and non-fictional language.

Fictional and non-fictional language have a certain semantic similarity, too, because, generally, both are meaningful (with the exception, perhaps, of works such as "Jabberwocky").[48] There are both important distinctions between syntax (e.g., grammar) and semantics (including meaning and truth) as well as important relations between them. It is possible to have a syntactically correct sentence that is not semantically meaningful ("Colorless green ideas sleep furiously" was Chomsky's famous example), and semantically meaningful sentences that are not true, of course (consider almost any sentence of fiction). The issue with LLMs is that they're very, even surprisingly, good at syntax and (perhaps) meaning, but not at truth.

How is the truth value of a statement, including the output of an LLM, determined? And who determines it? The reader of a text, the hearer of an utterance, or the judge in a Turing Test makes the determination by comparing the statement with known truths (making a correspondence between them) to see if they cohere. An AI could do the same, though LLMs currently do not.

But it's not only truth that we want. Philosophy papers aren't necessarily "true". Recipes may be "accurate", not "true". (Hence my preference for 'non-fictional' over 'factual'.) Determining whether a sentence is true (non-fictional) or false (fictional) requires something over and above mere linguistic ability, such as some kind of correspondence either with reality or with other linguistic data. (For just one example of the latter, without knowing which of two sentences S or $\neg S$ is true, a (consistent) system should be able to determine that at least one of them is false.) Current LLMs do not seem to have any mechanism for making such a determination, nor any "motivation" for doing so.

[47] Geoffrey Hinton argues that a better term is 'confabulate' (Rothman 2023; cf. Henriques 2024; Marcus 2024e,g.) Yet another term is 'bloviate' (von Hippel, 2023). And Weil 2023 suggests that LLMs are

> the Platonic ideal of the bullshitter, as philosopher Harry Frankfurt, author of *On Bullshit* [2005], defined the term. Bullshitters, Frankfurt argued, are worse than liars. They don't care whether something is true or false. They care only about rhetorical power—if a listener or reader is persuaded.

Of course, LLMs don't even care about that; they don't (yet) "care" about anything. Cf. Marcus 2024b. On the other hand, Buckner 2024, pp. 57–58, suggests that LLMs' "hallucinations" can be understood in a more positive light as a form of imagination; cf. Broad 2024.

[48] Though it's important to note that "Jabberwocky" can be *given* meaning, as Humpty Dumpty did (Carroll, 1871).

The other side of the coin is that humans are very good at *imputing* meaning where there isn't any that's obvious (several poems have been written in which Chomsky's sentence makes perfect sense).[49] Unfortunately, we're also very good at imputing *truth* where there isn't any; we're gullible.[50]

Is it possible to distinguish fictional from non-fictional language? There are certainly literary fictional linguistic styles that are clear (or nearly clear) marks of their fictionality, such as represented speech and thought (Banfield 1982; Rapaport et al. 1989, §3; Galbraith 1995; Segal 1995).[51] Of relevance to Turing's original Imitation Game, distinctions can also be detected between language use by males and females (Argamon et al., 2003). And:

> Previous studies [Zhou et al. 2023] show that human and machine-generated misinformation has distinctively different linguistic features that can aid in identifying misinformation. (Kabir et al., 2023, §2, p. 2, col. 2)

The question is whether these distinctions would be easy enough to detect in all or most cases in order to obviate any mistaken identity. In some circumstances, it may not matter if language is produced by a man or a woman, or even by a human or a machine. But distinguishing between truth and falsity is, of course, an important ability that involves critical thinking skills.[52]

3.8 Reactions to the Output

We don't say whether a *human* is Intelligent based on how his or her neurons fire; we make that judgment based on (our responses to) the human's output. Deter-

[49] You can read some of these at https://cse.buffalo.edu/~rapaport/675w/colorless.html

[50] On human gullibility, see O'Connor 2024, §3, p. 12, and this "Blondie" comic strip from 26 February 1982:

> Dagwood, reading newspaper: "Oh, no!"
> Blondie: "What's the matter?"
> Dagwood: "This article says that people who believe what they read are fools."
> Blondie: "So?"
> Dagwood: "I believed it!"

[51] However, there are grammatical similarities between represented speech and thought, on the fiction side, and the phenomena of quasi-indexicality, logophoric pronouns, and *de se* beliefs, on the non-fictional side. See Castañeda 1966, 1967; Perry 1979; Sells 1987; Rapaport et al. 1997.

[52] Should LLMs be blamed or held responsible for not paying attention to truth? Probably not the current ones; after all, they weren't designed for that! But paying such attention is needed for "real" Intelligence; see §6, point 5, below. Interestingly, critical thinking should be required of two participants in Turing Tests: The computer needs to be able to think critically about its training data (e.g., to avoid bias or trivial errors), and the human judge needs to be able to think critically about the computer's output. For more on fictional vs. non-fictional language, see Macdonald and Scriven 1954.

mining whether a Turing Test is passed is entirely in the purview of the judge. Neither the subject being tested nor even the computer's program matters. *How* a system passes—whether by human-like cognition or donkey work—is irrelevant. The judgment should *not* be made on the basis of *how* (because if it *is* donkey work, then the judge might say 'no'). The "outer" is more relevant to the judgment than the "inner".[53]

Because human judges ignore the internal processing and because of the relative indistinguishability of fictional from non-fictional language, the output of LLMs is taken at face value. This is . . .

> . . . another reason that a Turing test will eventually be passed. It is less interesting from a computational point of view, more so from a sociological point of view. It is simply that—to return to the earlier discussion of the Internet dog—for whatever reasons (and what these are is worth exploring), *humans tend to treat other entities with which they interact as if they were human.* (Rapaport, 2000, §9, p. 486, italics added)

Here are three aspects of this phenomenon.

3.8.1 The Eliza Effect

> [T]he urge to personify these systems is, for many people, irresistible, an extension of the same impulse that makes [us] see a face on the Moon or attribut[e] agency and emotions to two triangles "chasing" each other around a screen.
> —Luccioni and Marcus (2023)

We almost cannot help but attribute causality, motive, etc., to inanimate objects. Consider animated cartoons and, especially, our almost universal and automatic reaction to Heider and Simmel's (1944) classic movie of triangles and circles that we cannot help but see as cognitive agents. It is difficult, if not impossible, to react to them or describe them "neutrally".

This phenomenon—perhaps related to pereidolia[54]—has been called the "Eliza Effect", having famously been seen when some people reacted to the Eliza computer program as if it were a real psychotherapist and not merely a computer program.[55] Such people have, it's been said, unintentionally played the role of judge in a Turing Test and have accepted Eliza as passing it. If, for example, having a

[53] Cf. Stanley Cavell's views, discussed in O'Connor 2024, §3, p. 11.

[54] https://en.wikipedia.org/wiki/Pareidolia

[55] On Eliza, see Weizenbaum 1966, 1967, 1976. On the Eliza Effect, see https://en.wikipedia.org/wiki/ELIZA_effect

"purpose" or a "communicative intent" is necessary for language use (cf. Bender and Koller 2020, p. 5187, col. 1), then the Eliza Effect would be that the audience *assumes* or abductively (hence defeasibly) *infers* that the producer has such a purpose or intent. In the Eliza Effect, the human participants determine, unintentionally, that what they have been interacting with exhibits Intelligence.[56]

3.8.2 Response Dependence

> The meaning of information is given by the process that interprets it.
> —Edward Fredkin (10 June 2009), https://x.com/edfredkin/status/2112341253

The Eliza Effect is related to Diane Proudfoot's "response-dependent" interpretation of the Turing Test:

> [W]hether or not machines think is *in part* determined by social environment, in the form of the interrogator's responses (Proudfoot, 2005, my italics)

> [A]n entity is said to be intelligent (or to think) *only if* we respond to it in a certain way (Proudfoot, 2013, p. 397, my italics)[57]

Where the Eliza Effect does not occur on purpose (it is unintentional), response dependence can be more purposeful.

Is response supposed to be merely a *necessary* condition of Intelligence, as Proudfoot literally says? If so, it would require a prior characterization of Intelligence. But it is response as a *sufficient* condition that Turing seems to have had in mind, especially given his claims about "the use of words and general educated opinion". Intelligence is in the eye of the beholding judge; it is not necessarily something internal to the system being tested. Thinking that an LLM is Intelligent tells us more about *us* than it does about either the LLM or Intelligence. Sejnowski (2023, §4, p. 317) makes a similar point when he comments that "LLMs may effectively be carrying out a ... reverse Turing test, one that tests the intelligence of our prompts and dialog by mirroring it back to us." (We see what we want to see, as in Harry Potter's Mirror of Erised.)

[56] On the related notion that "comprehension automatically implies belief" and that such (temporary) belief is distinct from, and precedes, understanding, see Pennycook et al. 2012. Related to these ideas are "the mechanisms of a psychic's con" (Bjarnason, 2023).

[57] See also Proudfoot 2017, §13.2, and cf. Schank nd, p. 14: "the failure of experts to distinguish between imitations and the real thing should not be taken as much more than a statement of the competence of the experts." An earlier argument in favor of the prominence of the judge's response is in Rapaport 2000, §6.1; Rapaport 2023, §18.8.4.

In 1988, I had a conversation with the psychotherapist version of Eliza in which I pretended to be Hamlet.[58] The interaction was predictable Eliza: repetition, pattern matching, canned responses. But at one point towards the end of the conversation, the Eliza program repeated a comment from earlier in the conversation that, when placed in conjunction with the conversation at the end, caused me to remark "How interesting! I never thought of that!". A random conjunction of two sentences caused me to think about the play *Hamlet* in a new way. But, given the nature of Eliza, that I was impressed said more about me than about Eliza.[59]

3.8.3 The Intentional Stance

Responding to a system as if it were Intelligent is to take Dennett's Intentional Stance towards it. This is the "stance" one takes towards an "Intentional system", i.e., one

> whose behavior can be (at least sometimes) explained and predicted by relying on ascriptions to the system of beliefs and desires (and hopes, fears, intentions, hunches, . . .). (Dennett, 1971, p. 87)[60]

Whether or not the entity is "really" Intelligent, "*Treating* them as intelligent agents" (Luccioni and Marcus, 2023, my italics) is to take the Intentional Stance towards them. Given Dennett's (2023b) admonitions not to trust counterfeit people, this would seem to be an untoward consequence of his theory.

The Intentional Stance may need to be supplemented. In the quotation from Proudfoot, I italicized "in part": Proudfoot's response-dependence theory seems to allow for there to be *other* aspects of Intelligence besides our response. The intuitions at play here suggest a distinction between (a) the external output of a putatively Intelligent system to which we might take the Intentional Stance and (b) the internal processing that produces that output. We can then ask questions such as: Are there features of the internal processing that are *directly* responsible for such external Intelligence? (I.e., is the *processing* "Intelligent"?) Can there be one without the other?[61] Which is "more important"?

[58] The interaction is online at https://cse.buffalo.edu/~rapaport/hamlet.script.html

[59] The other side of the response-dependence coin is the first-person point of view. For the first-person perspective in (GOF)AI, see Rapaport 2023, §18.8.4, and the citations at https://tinyurl.com/rapaport2023-1884.

[60] I will assume the reader's familiarity with Dennett's theory of the Intentional, Design, and Physical Stances. See Rapaport 2023, §12.4.1, p. 290, for discussion. The Intentional Stance may be the default mode of understanding.

[61] Consider this "tweet" by Jaron Myers on 25 January 2022: "I've seen too many youth pastors be like 'Be careful on TikTok, it's just girls dancing in swimsuits' and I'm like bro. . . it's an algorithm". The dancers are the output of the algorithm; more precisely, they are pixels interpreted by viewers; there are no dancers in the algorithm.

3.8.4 Summary

Non-expert judges (and some expert ones) see Intelligence in LLMs because of these phenomena. Expert judges who do not see Intelligence in them expect more (see §6). In none of these cases does it follow logically that the entity is really Intelligent (if there is such a thing). "[T]he mere *subjective appearance* of meaning is not a reliable indicator of a **system's** understanding" (Michael, 2020, original italics, my boldface). That *we see* meaning in a text does not imply that *there is* meaning in the text or that the producer of the text put meaning into it. Yet there might be.[62]

There are several possibilities:

1. The entity in its internal workings might not be Intelligent at all, even if its program is based on (GOF)AI. This is arguably the case with the Eliza program.

2. The entity's internal workings might not *seem* to be Intelligent, yet it might after all come to satisfy all of the requirements of Intelligence (whatever they may be). This would be the situation with a program that behaves by "donkey work". It is consistent with the view of AI as engineering, and it may be the situation with LLMs.

3. The entity might really be Intelligent in its internal workings. It might be the case that any system capable of being responded to *as if* it were Intelligent must have something in its workings that produces that kind of response in us. This is (was?) the strategy of cognitively-inspired GOFAI, as urged most famously by Simon, and it may be what Buckner 2024 has in mind.

Note, by the way, that it also does not follow logically that the human (intentionally or not) will *judge* a *really* Intelligent entity *as* Intelligent.

The most serious consequence of LLMs is that, even though they may not really be Intelligent, and may be imperfect in many respects, they are convincing enough to lead uncritical (and even some highly critical) people to accept them at face value. They may be more than convincing, in fact, given our tendency to impose interpretations on the most minimal syntax, as in the triangles case cited above.

[62] Compare Intelligence as a response-dependent concept with color as a secondary quality (Pettit, 1991, p. 601). Is a "red" object "really" red? Not if redness is solely a secondary quality. But there must be something "in" the object that causes us to perceive it as red (that causes a red percept in us or that causes us to judge it as red). That "something" is not redness itself, but a cause (in the right circumstances) of a red appearance to us. Likewise for Intelligence: There must be something "in" the system that causes us to judge it as Intelligent. That something is not (or need not be) Intelligence itself; it could be donkey work.

4 LLMs and the Turing Test

So, LLMs are statistical, linguistic data processing systems that output fluent and apparently meaningful language. They are precisely the kind of machines that are the focus of the Turing Test. Do they pass it? Or are they Chinese Rooms?

4.1 Do LLMs Pass a Turing Test?

You can't say that a system has or has not passed "the" Turing Test, because there is no such thing. There are only various Turing(-like) "tests", given by different judges, not all of which are intended as tests (or even "games")—witness the Eliza Effect. Some systems pass some of them, while others don't. Do LLMs pass any Turing-like Tests? Probably the most accurate—if unhelpful—answer is that sometimes some do,[63] and sometimes some don't.

LLM computations certainly can produce output that is indistinguishable from human intellectual capacities, thus being able to be "regularly mistaken for a human being" (recall §3.1), with all attendant ethical implications of that. Vasant Dhar (2024) observes that "For the first time, we can converse with an entity, however imperfectly, about anything, as we do with other humans." Or consider this claim:

> "The chat interface is the killer app," said Dr. Jonathan H. Chen, a physician and computer scientist at Stanford who was an author of the new study [Goh et al. 2024]. "We can pop a whole case into the computer," he said. *"Before a couple of years ago, computers did not understand language."* (Kolata, 2024, my italics)

I take this as partial evidence that they (seem to) pass a Turing Test. Further evidence, consistent with Turing 1950, is in their ability to write poetry.[64] I asked

[63] If so, they have done it less than 25 years after Turing's initial target date of 2000 (§6, p. 442) and well before his more cautious target date of 2052 (Turing et al., 1952). This is better than Simon and Newell's (1958) prediction that a computer would "be the world's chess champion" by 1968, something that did not happen until 29 years later. On that prediction, see "Digression on 'the End of the Century' " in Rapaport 2023, §18.3.2, p. 411.

[64] And to be "creative", in general; see, e.g., Raiola 2023; Chakraborty and Masud 2024. Note that creativity does not require being "brilliant"! During the recent Hollywood writer's strike, author and film director Nicholas Meyer posted a photo to Facebook showing an airplane trailing a sign: "Pay the writers, you AI-holes!" because one reason they were striking was that LLMs can write TV shows and movies *well enough*. It reminds me of Ralph Raimi's put-down of the humanities (uttered as a side comment during his mathematical analysis class at the University of Rochester, Fall 1966): Only scientists and mathematicians have the spark of creativity; people in the humanities just re-arrange words, something that LLMs are good at.

Google's Bard to "write me a sonnet on the subject of the Forth Bridge" (Turing, 1950, §2, p. 434). Turing's imagined digital computer said that it "never could write poetry" (p. 434), but Bard attempted to comply (23 March 2023).[65] And both ChatGPT and GPT-4 have written *poems* correctly proving that there are infinitely many primes (Bubeck et al. 2023, p. 5, Fig. 1.1; p. 6, Fig. 1.2. GPT-4 is a better poet!)

4.2 (Some) LLMs (Sometimes) Pass

There are claims that LLMs pass, but there are important caveats: In a somewhat infamous paper on "sparks of intelligence" in GPT-4, Bubeck et al. (2023, p. 4) claim that "a new LLM developed by OpenAI, which is an early and non-multimodal version of GPT-4 . . . , exhibits *many* traits of intelligence" (my italics). And Bommasani et al. (2022, p. 22) note that "*non-experts* have difficulty distinguishing *short-form* English text that was written by GPT-3 from that written by humans" (my italics). Note the qualifications of "many" (vs. "all"), "non-experts", and "short form".

Bubeck et al. (2023, p. 27) asked GPT-4 to execute some Python code:

> It is important to note that GPT-4 is not running the code on a Python interpreter, but rather simulating the code in natural language. *This requires a high level of understanding and reasoning about the code*, as well as the ability to communicate the results clearly. [my italics]

65

> Here is a sonnet on the subject of the Forth Bridge:
> The Forth Bridge, a marvel of engineering,
> Stands proud and tall across the Firth of Forth.
> Its soaring towers reach for the sky,
> A testament to human strength and skill.
> The bridge is made of steel and iron,
> And its construction was a feat of daring.
> It took years to build, and many lives
> Were lost in its construction.
> But the Forth Bridge is a symbol of hope,
> A reminder that even in the darkest of times,
> Humanity can achieve great things.
> It is a bridge to the future,
> A bridge to a better world.
> It is a bridge to hope.

Note, however, that although it has a sonnet's required 14 lines, it neither rhymes nor is in iambic pentameter. Turing's imagined computer seems to have been more honest about its abilities than Bard.

Does it "*require* a high level of understanding" (in order to produce such a result)? Or does it merely exhibit what Bubeck et al. *take* as a high level of understanding (without "having" such understanding as part of its processing)? Or does it show that only statistical prediction, *not* understanding, is required, as in the following comment?

> … thanks to decades of programming effort and hardware improvements [chess] programs running on smartphones can now beat human grandmasters. Yet the methods they use are alien to any processes that might plausibly be at work in human minds. *Rather than declare computers intellectually superior to us, we have collectively agreed that intelligence is not needed to play chess.* (Haigh, 2023, p. 37, col. 2, my italics)

Or, for that matter, does it show that such statistical prediction *is* understanding?

Bubeck et al. (2023) say that

> A primary goal of our experiments is to give a preliminary assessment of GPT-4's *intelligence* … [T]he evaluation of the capabilities and cognitive abilities of those new models have become much closer in essence to the task of evaluating those of a human rather than those of a narrow AI model. (p. 92)

Part of their evidence for this is that "GPT-4 successfully passes the classic Sally-Anne false-belief test from psychology" (Bubeck et al., 2023, p. 55, Fig. 6.1). But does it follow that GPT-4 has a more or less explicit theory of mind, demonstrating that Intelligent processing and responses can be gotten from neural-network processing? Or is a theory of mind merely implicit in the *output*? Or perhaps the evidence is flawed: "… doing psychology on large language models is harder than you might think" (Marcus and Davis, 2023).

Further suggestions that LLMs pass Turing-like Tests are the many examples in the literature of LLMs passing various professional examinations in medicine, law, etc. (Kung et al., 2023; Katz et al., 2024). But the reality might be different:

> Human tests are designed using psychometrics, most often Item Response Theory (IRT). … IRT yields no guarantee that this validity is true for non-human test takers, such as test takers that are AI algorithms or aliens from another planet. Because AI models answer human test items in different ways than human test takers do, we cannot assume a high test score means a smarter AI model. The IRT model was never given the data that would be necessary to make a reliable discrimination among smart or shallow AI models. (Roschelle, 2023)

On this view, LLMs have *not* passed those tests. Yet they *seem* to have passed them according to Turing's standard.

There are also examples of LLMs failing Turing Tests. In evaluating ChatGPT's vs. Stack Overflow's answers to software engineering questions, Kabir et al. (2023, §7, p. 9) say this:

> It is evident from our results [that] ChatGPT produces incorrect answers more than half of the time. 54% of the time the errors are made due to ChatGPT not understanding the concept of the questions. Even when it can understand the question, it fails to show an understanding of how to solve the problem.

This could be viewed as evidence that ChatGPT has *not* passed this version of a Turing Test. Must a putative AI always pass all Turing Tests in order to be considered Intelligent? If a human passes a medical exam but does not know how to treat a disease—to apply that knowledge—would we say that the human was not Intelligent, or just a bad doctor?

So, even if *some* LLMs *sometimes* are judged (sometimes by non-experts) to have passed *some* Turing Tests, it is not at all obvious that they have joined the ranks of the Intelligent.

4.3 Internal Processing? Or External Output?—Part II

> Often they [the aliens] drew out of him indirectly much more knowledge than he consciously possessed . . . .
> —C.S. Lewis, *Out of the Silent Planet* (1938, Ch. 16, p. 102)

As we saw in §3.6, two ways (or levels) of looking at a putative Intelligent system are its internal (or lower-level) mechanisms and its external (or higher-level) output.

An LLM's internal mechanisms are its statistical algorithms and training that result in next-token prediction (the Design and Physical Stances). As with Leibniz's (1714) Mill, it is hard to see them as Intelligent when viewed from those lower-level stances (especially the Physical Stance). And its external output is the sometimes astonishing, sometimes embarrassing generation of grammatical and apparently meaningful language, open to a higher-level, "psychological" explanation or description, treating them as Intentional Systems. Compare this to humans. A human's relevant internal mechanisms are its patterns of neuron firings. And a human's external output is its (sometimes embarrassing) generation of grammatical and meaningful language.

"Mental illnesses" or "psychological differences" in humans are produced from lower-level neuron firings that differ from the "norm". Consider how psychiatrists or clinical psychologists deal with these. In some, and conceivably all, cases, treatment with medicines that physically, chemically, or biologically alter the neuron firings will "cure" the illness or modulate the differences. This deals with issues at a "lower" level (as in the Physical and Design Stances).[66] But in some, if not all, cases, "talking cures"—things like psychoanalysis or cognitive behavioral therapy—that operate at the psychological level can also change the external output. This deals with behavior at a "higher" level, primarily from the Intentional Stance (though there might be some attendant, and more permanent, changes at the lower levels).

Should one of these stances take precedence? Eliminative materialists certainly think that the Physical (or possibly the Design) Stance should:

> Matti Tedre suggests the best way to curb sensational claims about LLMs is to replace "LLM" with "statistical model of language." Then the extinction prophesy [that humans go extinct from our inability to control AIs that are powered by LLMs] becomes: "humans go extinct from our inability to control statistical models of language." (Denning, 2023b, p. 24, footnote c)[67]

Perhaps. But, if we wanted to say that humans might go extinct because of our inability to control our political and social behavior, does replacing this with "humans will go extinct from our inability to control *our neuron firings*" help? This is a situation where taking the Intentional Stance *rather than* the Physical or Design Stance is important.

Turing Tests are judged solely on responses to external output. The internal mechanisms are irrelevant to that judgment. It is of some interest, of course, how (the appearance of) Intelligence is produced from next-token-prediction algorithms as well as from neuron firings. But whether there must be some Intelligence-producing structures at that lower level is a secondary, perhaps empirical, question. For example, suppose that an LLM gives the wrong answer to a subtle linguistic test such as the Winograd schema[68] or an interpretation of a *de re* vs. a *de dicto* context.[69]. What that shows is that the sort of statistical, deep-learning techniques

[66]Godfrey-Smith (2024, p. 34) warns of the moral problems of "invasive neuroscientific work" involved in studying animal cognition. Thus, we might have to resort to more of what I am calling "psychological" methods, including first-person (heterophenomenological?) descriptions or judges' responses to behavior, as Turing advocated. Still, we need to have high standards for ascribing Intelligence or consciousness; see §6, below.

[67]Cf. a similar recommendation from the Churchlands in MacFarquhar 2007.

[68]Levesque 2017; Kocijan et al. 2023.

[69]Zhang and Davidson 2024. For background, see Castañeda 1967; Rapaport et al. 1997.

that it is based on—its stochastic parroting or donkey work—are not capable of dealing with such issues. And, insofar as those issues are considered important as a part of Intelligence, that points to a limitation of the technique and the need for other techniques to be added to it.[70]

Thus, judges are basically forced to treat the systems from the Intentional Stance, and they are prone to Eliza Effects. As a consequence, credulous judges (mostly, but not exclusively, non-experts) are overly quick to give a passing score to AI systems whose output is apparently indistinguishable from humans. Incredulous judges (experts) will require higher standards.

The difficulty in identifying the output as coming from an AI system is exactly what Turing predicted when he said that the use of words would alter. If we judges cannot distinguish an AI system's output from that of a human, we will inevitably treat the two kinds of Intelligent systems alike, leading to various social, ethical, and legal problems.

If their output is really indistinguishable from that of humans, we need to be (reasonably) skeptical and to think critically about the output. Of course, the same goes for output generated by humans. A consequence of the Turing Test and the ability of LLMs to (seem to) pass them is that it is more important to treat all output skeptically and critically than it is to know who (or what) generated the output.

Here is another way to think about this. Consider two more-or-less parallel situations:

1. Human cognition is nothing but neuron firings.
   Cognitive behavior is epiphenomenal.

2. LLM "cognition" is nothing but statistical prediction.
   The output of an LLM is epiphenomenal (besides being untrustworthy).

Now consider two possible reactions to situation 1 ...

a The only thing that has to be studied is the neuron firings.

b On the other hand, everyone in their normal, day-to-day activities pays more attention to the cognitive behavior.[71]

... and to situation 2:

a The only thing that has to be studied is the statistical prediction.

[70] It is, of course, possible that scaling up the old techniques might address the problem. I doubt that (as do others, e.g., Marcus 2024h), but I won't rule it out here. Here, I'm merely pointing to aspects of Intelligence that need to be addressed.

[71] For a humorous take on this, see "Elementary Physics Paths", https://xkcd.com/2933/

b But Turing Test judges (both expert and non-expert) pay more attention to the output.

If an entity is to be taken as Intelligent, it is (b) that matters (be it an LLM, extraterrestrial, non-human animal, or human). The underlying processes are irrelevant (be they neuron firings, statistical machine learning, or donkey work of any kind).[72] Therefore, the Intentional Stance/response-dependent approach is the relevant one. The important features are the output and its interpretation by us, not the underlying processes.

GOFAI tries to computationally implement the cognitive behavior, so that the *program* at the Design level would embody the Intelligence, and so that the cognitive output and the internal processing would mirror each other. Stich and Ravenscroft (1994, pp. 14–15) make a similar distinction between "external" folk psychology as judged by observers and "internal" folk psychology that "is part of the mechanism subserving" those features.[73] The Churchlands, famously, deny the importance of external folk psychology (and hence the usefulness of any kind of internal folk psychology). As Colin McGinn (1999) puts it, their

> position is that folk psychology was cobbled together in an earlier, pre-scientific age, as a speculative theory of what causes people's behavior, and it is high time to examine it critically with a view to finding a more streamlined theory of our inner workings.

By analogy, we should ignore an LLM's "psychology" in favor of its statistical machine-learning processing.

But surely, to whatever extent we humans have a folk psychology that the Intentional Stance capitalizes on (even if it is not a full-blown scientific theory), LLMs also have their own. In our ordinary, everyday dealings with an LLM, we have to treat it via its (folk) psychology, i.e., from an Intentional Stance. We can't very well manipulate its statistical algorithms in ordinary interactions with it, any more than we can manipulate a human interlocutor's neuron firings.

Trott et al. (2023) and Han et al. (2024) offer such "psychological" studies of LLMs. These are high-level, behavioristic studies of the output of LLMs, comparing it to the psychology, hence high-level mental—as opposed to neurological—behavior of humans. However, there is a problem with dealing with LLMs only "psychologically": One aspect of Turing's presentation of the Turing Test is that you should be able to ask the system why it said something:

[72]Irrelevance has to be taken with a grain of salt: The field of AI conceived as computational cognition requires that the processes be at least algorithmic.

[73]Proudfoot 2005 makes a similar distinction between "internalist" and "externalist" interpretations of the Turing Test.

> Interrogator: In the first line of your sonnet which reads 'Shall I compare thee to a summer's day', would not 'a spring day' do as well or better?
> Witness: It wouldn't scan. (Turing, 1950, §6(4), p. 446)

Turing's point is that we would accept that response as more than mere "artificial signaling" (Turing 1950, §6(4), p. 446, quoting Jefferson 1949, p. 1110, col. 1). But should we?

> It's also silly to ask these sorts of models for questions about themselves; ... they don't actually "know"; they are just generating different word strings on different occasions, with no guarantee of anything. (Luccioni and Marcus, 2023)

Lance Fortnow (2023) actually tried this with ChatGPT, in the context of the "black box" problem for neural networks ("seek[ing] ... algorithms that can explain themselves"):

> You can speculate why Google Translate uses male pronouns but you can't ask it why. But you can ask ChatGPT.
>
>> I used female pronouns in the translation because Vera T. Sós is a female name. Vera is a feminine given name, and since the context of the original text is about her life and accomplishments, it is appropriate to use female pronouns in the English translation to maintain clarity and accurately convey information about her.

In the comments section of his blog, I asked Fortnow, "How do you know that ChatGPT's 'explanation' is really that, rather than merely more statistical language parroting?". He replied as follows:

> There is some circular reasoning here and the best you could hope for is an overly simplified view of its reasoning. When I have been asking ChatGPT to explain its actions, it does seem to be making a good effort at it.

So Fortnow seems to side with Turing.[74]

But is that kind of "higher-level explanation" really what is wanted (especially in the black-box problem)? Or is it "merely" an after-the-fact rationalization as in the "explanations" that some split-brain patients give for their odd behavior?[75] Perhaps such a rationalization is, in fact, a good high-level summary of the low-level statistical processing. Still, is such an explanation trustworthy? Bommasani et al.

[74] Johnson-Laird and Ragni 2023 offers a possible way around this problem. Cf. Kolata 2024: "The computer may not be able to give a clear explanation of its decision pathway, but does that matter if it gets the diagnosis right?"

[75] Gazzaniga 1989, esp. p. 951; Wolman 2012, esp. p. 262, col. 2.

(2022, p. 126) note that "It is important to be discerning of the difference between the ability of a model to create plausible-sounding explanations and providing true insights into its behavior." Chalmers (2023a) observes that "Current LLMs seem to have especially limited self models: that is, their models of their own processing and reasoning are poor." The same can be said about us! Perhaps all of our own explanations of our behavior are similarly illusions or rationalizations. So relying solely on psychology is not the answer, either.

To ask what an LLM does and to answer that it predicts the next word (as Shanahan (2024) recommends) is like asking what I'm doing now and answering "firing neurons". That's true, but I'm also writing an essay (*by* firing neurons).[76] We need to do "psychological"/Intentional-Stance analyses of LLMs *in addition to* statistical/"neural"/Physical-and-Design-Stance analyses. Although an LLM's explanations of its behavior may not be the real reasons it behaves as it does (where the real reasons are its statistical predictions), is this any different from our own psychological explanations of our behavior (where our real reasons are neural)?

5 If LLMs Pass, Are They Intelligent?

5.1 Introduction

If an LLM is *taken as* Intelligent by a judge (i.e., is *judged* to have passed a Turing Test), *is* it Intelligent?[77] That, of course, is precisely the question that Turing "believe[d] to be too meaningless to deserve discussion" (§6, p. 442). He did not want to draw an ontological inference from an epistemological premise. Turing would not say that LLMs *can* "think" or *be* Intelligent, only that they would pass a Turing Test, period, full stop.

He went on to say that eventually we would *call* their output 'thinking' (or 'Intelligence'): "the use of words . . . will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (§6, p. 442). As Lakoff and Johnson (1980) showed us, words can alter by metaphorical extension (for discussion, see Rapaport 2000; Rapaport 2023, Ch. 18). But another reason in the present case may be that eventually we will simply no longer *care* whether an AI is "really" Intelligent; we'll just become habituated to treating them so: As Martin Davis (2017, p. 157) points out,

> . . . no one seems to hesitate to speak of a "smart" phone; one doesn't hear:

[76] Cf. Rapaport 2023, §16.9, esp. p. 379.

[77] Or is it that such LLMs "appear to be able to fool the Turing Test" (Savage, 2024a)? Is it even possible to fool the Turing Test? Not if passing is definitional of—or a replacement for the idea of—thinking.

> "well of course the phone isn't really smart in the way people are". One can predict that if and when computers have attained the ability to do all of the things that are thought to constitute intelligence in people, people will not hesitate to use the word "intelligence" in referring to that behavior.

Turing also went one step further: "general educated opinion" will *also* "have altered"; i.e., we will have changed our understanding of both machines and thinking.

Have those two things happened? It's safe to say that most researchers—and even some ordinary people—no longer think that "machines can think" is an oxymoron, i.e., that computers can't think *simply because* they are machines. The issues now are: (A) Have they passed a Turing Test? (B) If they have passed, have they passed in the "right" (e.g., cognitive) way? And, of course, we cannot help but ask (C) *Are* they Intelligent? But, if they have not (yet) passed, is that because we haven't yet gotten "real" artificial (general) intelligence? Or is it because real artificial general intelligence is logically or practically impossible?[78]

Many non-expert judges—and even some AI experts (Bubeck et al., 2023)—have judged LLMs as passing and therefore Intelligent. Searle famously claimed that passing is *not* a sign of Intelligence. And some AI researchers—though not all—are still hesitant to call what the LLMs do 'thinking'. Why? Let's see what Turing had to say.

5.2 Turing's Replies to Objections

Turing (1950) considered various reasons for rejecting his test (or denying Intelligence to a passing entity). Some of them we can dismiss, such as the Theological Objection (§6(1), p. 443) and the Argument from Extra-Sensory Perception (§6(9), p. 453). We can also dismiss the Heads in the Sand Objection (§6(2), p. 444) that "the consequences of machine thinking" are "too dreadful" to think about (an early statement of the Singularity). However, even if not an objection, those potential (or actual) dreadful consequences do need to be thought about. The Mathematical Objection from Gödel's incompleteness theorem (§6(3), p. 444) has engendered much discussion, but it seems irrelevant to the LLM issue (see Rapaport 2024b for an overview). The present issue is only whether LLMs are as Intelligent as an ordinary human. Does an AGI need to be able to do math? We certainly expect most computers to be able to do math. But Turing did not think that math skills were necessary for passing a Turing Test. Even Turing noted that giving wrong answers

[78]On logical impossibility, see Landgrebe and Smith 2023; Rapaport 2024a. On practical impossibility, see Marcus 2024j; Rapaport 2024b. (And on the practical impossibility—better: unlikelihood—of a *musical* Turing Test, see Neely 2024; thanks to Michael I. Rapaport for the pointer.)

to mathematical questions is *not* evidence of non-intelligence. (But see Mirzadeh et al. 2024.)

5.2.1 The Argument from Consciousness

The Argument from Consciousness (§6(4), p. 445) is a bit more applicable. LLMs *can* write poetry and compose music (Bubeck et al., 2023). But, importantly, LLMs are not thus artistic "*because* of thoughts and emotions felt" but indeed only "by the chance fall of symbols" (p. 445, my italics).

Does Intelligence require "felt" thoughts and emotions? Can it emerge from the chance fall of symbols? Will it be (merely?) a "polite convention" (p. 446) to say that it is Intelligent?[79] Turing "think[s] that most of those who support the argument from consciousness could be persuaded to abandon it" (p. 447), but the Google employee who was fired because he believed that Google's LaMDA LLM was sentient did accept it.[80] Indeed, both the current reaction of many experts to LLMs as well as Turing's reply to the argument from consciousness suggest that real experience (whatever that might mean) is necessary for Intelligence, not mere reading knowledge in the sense of statistical training on how to predict the next character. (We'll return to this in §6, point 11.)

5.2.2 Arguments from Various Disabilities

Some of the Arguments from Various Disabilities (§6(5), pp. 447ff) may be "frivolous" (p. 448), such as "the inability to enjoy strawberries and cream".[81] Such enjoyment may not be relevant to Intelligence (especially as functionally defined), but interaction of some kind with the physical world may well be needed. After all, some LLMs are supposed to be multimodal. We'll return to this in §5.3.

The issue of machines making mistakes (pp. 448f) is a curious one. As Turing suggests, people make them, too, so why shouldn't computers? But the experiences with LLMs suggest a slightly different response. They obviously do something along the lines of making mistakes when they "hallucinate" (let alone "deliberately introduce mistakes in a manner calculated to confuse the interrogator" (p. 448)). Would citing the sources for assertions that an LLM makes solve this problem (as suggested by van Dis et al. (2023, p. 228, col. 1))? Perhaps, but only if the sources cited are real! The currently most infamous example of this is that of the lawyer who relied on ChatGPT to write a legal brief for him, which ChatGPT readily

[79] Another alternative—that Intelligence can be *perceived in* "the chance fall of symbols"—was discussed in §3.8.2.

[80] https://tinyurl.com/brodkin2022

[81] On this, see Long 2023.

did, making everything up, even insisting that "the … cases I provided are real" (Weiser, 2023a,b; Weiser and Schweber, 2023).

On the other hand, talk of hallucinations, etc., "attribute[s] unearned agency to the machine, when in fact the "made up stuff" is simply statistical inference from the training data" (Denning, 2023b, p. 25, cols. 1–2). This raises the question of when we should use the Design Stance (it's "simply statistical inference") rather than the Intentional Stance ("attribute … agency")! That, in its extreme form, is one of the reasons that the Intelligence of LLMs is worrisome. As Turing noted, "Errors of conclusion can only arise when some meaning is attached to the output signals from the machine" (§6(5), p. 449). Such meaning must be "attached" *by the user*, i.e., is attributed to, hence is external to, the computer. This is consistent with the response-dependent interpretation. It also suggests that the computer itself does not self-attribute meaning.

But we should also keep in mind Sejnowski's (2023) "Parable of the Talking Dog" who can speak but makes things up. Should we be disappointed by the latter? Or amazed by the former? We may be disappointed by the tendency of LLMs to hallucinate, but we should be amazed by the fact "they are superhuman in their ability to extract information from the world's database of text" (Sejnowski, 2023, p. 311).

5.2.3 Lady Lovelace's Objection

Ada Lovelace's positive statement is the universal one that a computer "can do whatever we *know how to order it* to perform."[82] But the (in)famous Lady Lovelace's Objection (§6(6), pp. 450ff) is the negative uniqueness statement that a computer "can [*only*] do whatever we know how to order it to perform".[83] Turing's reply is, roughly, that computers could be programmed to learn new things, hence to do things that they were not explicitly programmed ("ordered") to do. LLMs *can* clearly do things that they were not explicitly programmed to do, such as pass bar association and medical school tests. So Turing was correct in his reply to the Lovelace Objection, but possibly for the wrong reason: These LLMs didn't explicitly *learn* law or medicine, but they statistically generated answers based on the textual databases that they were trained on.

On the other hand, Melanie Mitchell (2023) notes that

> Although we assume that humans taking a standardized test have not already seen the questions and answers, the same is not necessarily true for a large-

[82] Menabrea and Lovelace 1843, p. 722, Note G; italics in original.

[83] After correctly quoting it on p. 450, Turing (intentionally?) misstates it on p. 454: "the machine can *only* do what what we tell it to do" (my italics), but then corrects(?) himself in a footnote on p. 459: "Compare Lady Lovelace's statement (p. 450), which does not contain the word 'only'."

> scale AI system like GPT-4, which has been trained on vast swaths of digital media, some of which may have included the questions GPT-4 was later tested on.

But how does that differ from a human who studies for an exam and perhaps does test preparation using guides that include old exams? (When I was in high school, the SAT re-used an old exam that students had practiced on, and almost everyone in the class got a perfect score!)

A better test is whether the knowledge (or "knowledge") that these systems (and these people!) have can be *applied*. And the evidence for that is currently underwhelming: Did the test-passing LLMs implicitly learn law or medicine? Arguably not: Passing these tests is one thing. Being able to apply their ostensive "knowledge" may be quite another thing (see §1.1, esp. footnote 12, above).

5.2.4 Argument from Informality of Behaviour

According to the Argument from Informality of Behaviour (§6(8), p. 452), "It is not possible to produce a set of rules purporting to describe what a [hu]man should do in every conceivable set of circumstances" (cf. Landgrebe and Smith 2023). Hence, common sense is needed, and indeed this seems to be something that LLMs lack.[84] Turing agreed that computers (at the time) lacked common sense but rejected the conclusion that therefore "we cannot be machines" (p. 452). I think we can interpret this as Turing's rejection of the claim that therefore human-level Intelligence is not computable.

5.3 The Chinese Room Argument

> How can I speak a language without knowing I can speak it?
> —J.K. Rowling, *Harry Potter and the Chamber of Secrets* (1999, p. 196)

5.3.1 Introduction

There is one more objection, not considered by Turing, namely Searle's argument against Claim 1 (§1.1, above) that there is more to Intelligence than passing. Could

[84]For discussion of this point and the need for GOFAI in general, see: Levesque 2017; Boden 2018, pp. 86–89; Garnelo and Shanahan 2019; Marcus and Davis 2019; Seabrook 2019; B.C. Smith 2019; Marcus 2020, 2023c, 2024a; Landgrebe and Smith 2021; Sablé-Meyer et al. 2021; Brachman and Levesque 2022, pp. 48f; Quilty-Dunn et al. 2023; Wolfram 2023, pp. 79ff. The precise characterization of "common sense" may not be clear; cf. Buckner 2024.

the judge in a Turing Test be mistaken?[85] Is passing a Turing Test merely circumstantial evidence?

Searle's (1980) Chinese Room Argument can be seen as a counterexample to the Turing Test, arguing that a Turing-Test-passing entity need not be Intelligent.[86] Turing envisaged something like the Chinese Room set-up when he discussed the *viva voce* version of his game.[87] Assuming that an LLM passes a Turing Test, might it be a real-life Chinese Room?

The Chinese Room Argument consists of an argument from biology and an argument from semantics (Rapaport 1986a, 2000; Rapaport 2023, Ch. 18). The biological argument (computer programs are non-biological; cognition is biological; hence, no non-biological computer program can exhibit cognition) is irrelevant to LLMs: No one has (yet) suggested that LLMs are not Intelligent because not being biological; "thinking machine" is no longer an oxymoron. It is the semantic argument that is relevant.

5.3.2 The Semantic Argument

Applied to LLMs, the semantic argument is this:

(S1) An LLM is a purely syntactic program.
(S2) Cognition is semantic.
(S3) Syntax alone is not sufficient for semantics.
(S4) So, an LLM does not exhibit semantic cognition.

[85] Rapaport 2000, §6, argues against this possibility on the grounds that the judge's point of view takes precedence.

[86] Searle's 1980 essay does not talk about "intelligence" (except as that word is modified by 'artificial'). The abstract talks about "thinking". Searle talks most frequently about "understanding" and understanding language. And he treats such understanding as a special case of "intentionality".

[87] The logician Hartley Rogers, Jr., proposed a *viva voce* Chinese Room set-up long before Searle:

> Consider a box B inside of which we have a man L with a desk, pencils and paper. On one side B has two slots, marked *input* and *output*. If we write a number on paper and pass it through the input slot, L takes it and begins performing certain computations. If and when he finishes, he writes down a number obtained from the computation and passes it back to us through the output slot. . . . (Rogers, 1959, pp. 115, 117)

Note that Rogers said "a number", not necessarily "the correct" number! In the case of the Chinese Room Argument (as opposed to a Turing Machine), what corresponds to "correct"? Acceptable? Grammatical? Meaningful? Truthful? Compare the sentence 'Buffalo buffalo buffalo buffalo buffalo' (Rapaport, 2015), which is grammatical and meaningful, but not very acceptable (its truthfulness remains an empirical question). Note that Turing did not require the computer in the Turing Test to be accurate or truthful. In fact, he required it to lie if necessary in order to pass. But note that lying probably requires planning, something that LLMs (currently) lack (Bubeck et al., 2023).

Rapaport 1988 and 2000 argue (against S3) that syntax—the kind discussed in §2, above; not mere grammar—*can* suffice for the kind of semantics required for cognition. Do LLMs have that kind of syntax? Are their stochastic, deep-learning algorithms—which are purely syntactic—sufficient for Intelligence?

Searle (1982, p. 5) says that the Room's "formal program . . . attaches no meaning, interpretation, or content to any of the symbols." Does an LLM "attach" meaning, etc., to its symbols? Arguably, no. LLMs do not (yet)[88] have the kind of internal, "syntactic" semantics needed for full understanding. Although syntax does suffice for conceptual-role semantics, not all syntax is of the right kind. LLMs have some internal or syntactic semantics of the distributional sort, but they lack a full conceptual-role semantics (which might require GOFAI techniques). In Melanie Mitchell's words,

> These systems live in a world of language. . . . That world gives them some clues about what is true and what is not true, but the language they learn from is not grounded in reality. They do not necessarily know if what they are generating is true or false. (Quoted in Metz 2023b)

There are two lessons from this observation: First, "hallucinations" are not only based on the principle of garbage-in/garbage-out (which is also responsible for the bias problem),[89] but they are also an example of the fact that fictional and non-fictional language are similar enough that they can be confused (recall §3.7): "It is indifferent for a machine to refer to the real world rather than to talk about nothing or a possible but nonexistent world" (Perconti and Plebe, 2023, §4.2).

The second lesson is that to get the kind of referential semantics that Searle wants (what Harnad 1990 calls "symbol grounding"), what's needed—and what's missing from LLMs—are *internal representatives* of external objects (Rapaport 1995, 1998; Rapaport 2023, §18.10, p. 436). Distributional semantics and conceptual-role semantics without such internal representatives do not suffice for natural-language understanding. As Bisk et al. (2020, p. 8718) point out, "*You can't learn language from the radio.* . . . However, as a field we attempt this futility: trying to learn language from the *Internet* . . . ". And "these systems do not really understand anything. *The words refer to words, not to their meanings*" (Haikonon 2020, p. 76, my italics, as cited in Perconti and Plebe 2023, §4.1)

[88] Some versions of these programs are not "multimodal". But such multimodality will be needed for them to "attach" meaning to their linguistic symbols. Other versions, however, can begin to do this, e.g., when linked with visual programs such as DALL-E.

[89] Or "bias in, bias out" (Floridi, 2024). There can be non-garbage-related sources of bias, too; see Marcus 2024d. See also Waldo and Boussard 2024, p. 2: "the question to ask is not, 'Why do GPTs hallucinate?' but rather, 'Why do they get anything right at all?' "

The most recently famous version of this is Bender and Koller's (2020, p. 5185, col. 2) octopus "thought experiments illustrating the impossibility of learning meaning when it is not in the training signal". But this "impossibility" is false or misleading.

It is false because research on contextual vocabulary acquisition shows that meaning *can* be learned that way:[90] "A human may infer the meaning of a word when reading it for the first time in a book, on the basis of their prior understanding of the meaning of its linguistic context" (Michael, 2020).

And it is misleading because it depends on what is meant by "in the training signal": If the signal includes what humans have available to them in addition to *linguistic* "signal", then such a system can learn meaning. Bender and Koller (2020, p. 5188, col. 2) note that their octopus does not see what is being talked about "and thus would not be able to pick out the referent of a word when presented with a set of (physical) alternatives." It follows that, if it could see them, it might be able to understand.

They clarify this a bit later: "a system exposed only to form in its training cannot in principle learn meaning" (p. 5186, col. 2). By 'form', they mean what I call 'syntax', and they agree that meaning is a "relation between the form and something external to language" (p. 5187, col. 1). But syntactic semantics ("form") *can* suffice for understanding by "internalizing" the external things (Rapaport 2006, 2017; Rapaport 2023, Ch. 18). They seem to agree with this, too: "… if form is augmented with grounding data *of some kind*, then meaning can conceivably be learned to the extent that the communicative intent is represented in that data" (p. 5192, col. 2, my italics).

By contrast, the internal symbols that current LLMs deal with are not sufficient for cognition. At most, they include words (or phrases), but they don't include symbols for concepts or things, nor do they include symbols relating those words or phrases to the concepts or things that they name or describe. So they need more such symbols—the internal representatives (or "avatars") of external objects. And for this, they might need the ability to *perceive* things in the external world (Chalmers, 2023b) in order to get the internalizations.[91]

[90] Ehrlich 1995; Rapaport 2003b, 2005; Rapaport and Kibby 2007, 2014.

[91] Would they also need the ability to *manipulate* such external things? As Wittgenstein (1953, §2) suggested, to be able to get a block, you would have to know what a block is (conceptual-role semantics might suffice for this), you'd also have to know what it looks like (this might require referential semantics using an avatar of an external block), *and* then you would have to be able to use that information to find and get one. But that last step might not be necessary. If we communicate with an extra-terrestrial via Zoom or Facetime, we wouldn't be able, much less need, to use such physical manipulation to decide if it was Intelligent or could understand language.

Bommasani et al. (2022, p. 25) recognize this: "One salient difference between LLMs and human language acquisition is that human language is grounded to the real world". But this is merely a current practical difference, not a principled one. A multimodal LLM could be grounded:

> [W]e need not restrict the model to seeing only textual input. A foundation model might be trained on a wide range of different symbols: not just language but also computer code, database files, images, audio, and sensor readings. As long as it is just learning co-occurrence patterns of the sequences it is exposed to, then it counts as a foundation model by our definition. As part of this learning, the model might come to represent strong associations between a given piece of text and a particular sensor reading, or between a sequence of pixel values and a database entry. These associations might reflect important aspects of the world we inhabit and the language we use to talk about it. ... if the input symbol streams include diverse *digital traces of things in the world*—images, audio, sensors, etc.—then the co-occurrence patterns might contain enough information for the model to induce high-fidelity *proxies* for the required mapping. (Bommasani et al., 2022, pp. 48–50, my italics)[92]

The bottom line with respect to the Chinese Room Argument is this: In order for an entity to pass a Turing Test and *not* be a mere unIntelligent Chinese Room, items in the external world that ground internal words have to be internalized as "representatives", "avatars", "traces", or "proxies".

Moreover, the grounding data is inevitably going to be of the same nature as the linguistic form. If the external "meanings" are internalized, then words *can* refer to these "avatars", which are just other syntactic entities. Sejnowski (2023, p. 334) makes this clear:

> How could an "artificial cerebral cortex" be said to understand what a flower is if its entire universe consists only of disembodied language?[[93]] Keep in mind that by the time our brain receives sensory input, whether from sight, sound, touch, or anything else, it has been encoded in the activations of neurons. The activation patterns may vary by sense, but the brain's job is to correlate them all, using each input to fill in the blanks—in effect, predicting other inputs. That's how our brains make sense of a chaotic, fragmented stream of sensory impressions to create the grand illusion of a stable, detailed, and predictable world.

[92] Some earlier—GOFAI— work along these lines is the ability to align names with faces; see Srihari and Rapaport 1989, 1990; Srihari 1991a,b. See also Li et al. 2024.

[93] Cf. Jackson's Mary (1986).

Importantly, this is not the case just for LLMs: All of *our* language and internalized external entities are encoded in the "internal common currency" of neuron firings (spike trains) (Piccinini, 2016, p. 216).[94]

5.3.3 LLMs and the Chinese Room

> Searle's *reductio ad absurdum* argument is designed to show that the existence of a program *p* enabling fluent communication in [language] *l* is not sufficient to say that the system running *p understands l*.
> —Anders Søgaard (2023, p. 37)

In the Chinese Room, a human who knows no Chinese executes a computer program for processing Chinese, thereby passing a Turing Test judged by a native Chinese speaker. In a typical LLM system, a central processing unit (CPU) executes an LLM program for processing English, thereby (let's say) passing a Turing Test judged by a (not necessarily native) English speaker.

If the Chinese Room Argument is indeed a counterexample to Claim 1, who or what fails to understand Chinese? Searle took that entity to be the human in the Room, from which he inferred that no natural-language understanding was going on at all. He was correct that the person in the room did not understand Chinese. The person in the room is the analogue of the CPU of a computer executing an LLM program, and clearly CPUs just execute programs—any program: The very same CPU (of a Universal Turing Machine) that is given an LLM program will execute it and can be said to thereby speak Chinese to the exact same extent that it can do arithmetic or quantum mechanics or analyze the stock market when it is executing an arithmetical, quantum mechanical, or financial program. Whether it "understands" Chinese or arithmetic or whatever is a separate question. As Robin K. Hill (2016, §5) has noted, a program (an algorithm) can be considered in two different ways: An algorithm A can have the syntactic form "Do A" or the semantic (intentional, teleological) form "To accomplish goal G, do A" (Rapaport, 2023, Ch. 16). The CPU "does A" whether or not it "knows" or "understands" or is "aware" that it is accomplishing G.

However, from the native-speaker judge's point of view, understanding of Chinese is going on.[95] But, *pace* Searle (or Searle's interpretation), the human in the Chinese Room who does not understand Chinese is *not* the entity that passes the Turing Test without understanding Chinese. If it is not the person in the room, that

[94]Cf. Sloman 2010, whose notion of symbol "tethering" (as opposed to "grounding") seems similar to the role of internal representatives.

[95]Compare the reaction that has been attributed to Gassendi that all Descartes should have concluded from his *cogito* argument was that thinking is going on.

leaves two options: The natural-language understanding resides in the rule book or in the system consisting of the person in the room plus the rule book (the Systems Reply).[96] Searle is *correct* in saying that the human *alone* does not understand Chinese, because it is the human *together with* the rule book that understands (when the human executes the rules).

Although both of these options are on the right track, the situation is a bit more complex. In the case of an LLM, the rule book corresponds to the LLM algorithms. But there are other things to consider: There is the knowledge base (to use a GOFAI term) or the linguistic data (in neural-network terms) on which the LLM was trained. And although "a computer without a program is just a box with parts in it" (qFiasco, 2018, p. 38), and "without a computer, a program wouldn't be able to do anything" (Rapaport, 2023, §3.14, p. 67), it is also the case that without the CPU to execute the program—to bring it "alive"—both computer and program are inert. This is the difference between a static *program* and a dynamic *process*. So, the natural-language understanding resides in the dynamic execution, by the person-in-the-room CPU, of the rule-book algorithm, trained on linguistic data.

5.3.4 A "Whale Room" Argument

The Cetacean Translation Initiative project (CETI) uses LLMs to communicate with whales, who seem to communicate with each other using patterned sequences of clicks, called 'codas':

> In theory at least, what goes for English (and Chinese and French) also goes for sperm whale. Provided that a computer model can be trained on enough data, it should be able to master coda prediction. It could then—once again in theory—generate sequences of codas that a sperm whale would find convincing. The model wouldn't understand sperm whale-ese, but it could, in a manner of speaking, speak it. Call it ClickGPT. (Kolbert, 2023, p. 49, cols. 1–2)[97]

Why wouldn't it understand? For one thing, it would not have avatars of things that interest whales. Shafi Goldwasser, a computer scientist, is quoted by Kolbert:

> So you could say that's a goal for CETI—that you don't necessarily understand what the whales are saying, but that you could predict it with good success. And, therefore, you could maybe generate a conversation that would be

[96]The Systems Reply was anticipated, and attributed to Turing, in Mays 1952, p. 156: "What, on Turing's view, corresponds to the human mind is, however, not just the machine, but the machine plus the instructions fed into it."

[97]Also see the *New Yorker* cartoon by Ellie Black at https://www.newyorker.com/cartoon/a60538.

> understood by a *whale*, but maybe *you* don't understand it. So that's kind of a weird success. (p. 53, col. 1)

It's more than a "weird success"; it's almost a Chinese Room! But we need to be cautious here. Who or what might not understand the click conversation between whale and ClickGPT? Presumably the whale would understand it. If ClickGPT is really a Chinese Room, then Searle would certainly agreee that ClickGPT would not understand it. But we've just seen that a full syntactic semantics for clicks *would* enable it to understand the conversation. It is *we*, listening in to the conversation, who would not understand. If *we* play the CPU role in the Click Room, we would not understand, but the *room* would.

One difference between trying to understand a whale and the LLM situation is that whales are living creatures in a society of whales and other living creatures (Denning and Rousse, 2024). But LLMs are not (or not yet). By itself, this is not necessarily disqualifying. A recent TV ad for Etsy asks, "What does a robot know about love?", commenting that it takes a human to know. But is that really a limitation? What does a male human know about pregnancy or menopause? What does a white person know about being Black in the US? And so on. There are many things that we only "know" by description, not by "acquaintance" or direct experience. This is not typically taken as a limitation. (But cf. Jackson 1986.) Still, another reason that we would not understand "sperm whale-ese", even if we could communicate fluently with a whale, would be because the whale is a different "form of life" from us. If a whale could talk, we wouldn't be able to understand it (Wittgenstein, 1953, §327, p. 235e).

But there may be some hope. Shane Gero, lead biologist with CETI, said:

> For me, the most rewarding part about spending a lot of time in the culture of whales is finding these fundamental similarities, these fundamental patterns. ... And, you know, sure, they won't have a word for 'tree.' And there's some part of the sperm-whale experience that our primate brain just won't understand. But those things that we share must be fundamentally important to why we're here. (p. 47, col. 3–p. 48, col. 1)[98]

As Churchland (1984, p. 156,) notes, "we must not expect that the goals or concerns of an alien intelligent species will resemble our own, or even be intelligible to us." What does this have to say about whether an LLM with a full syntactic semantics could understand? After all, we have little have in common with LLMs (or they with us). All of this suggests that having an LLM-programmed robot that is raised in a human culture and programmed to learn as a child does (Turing, 1950; Chiang, 2019; Vong et al., 2024) will help with understanding on both of our parts.

[98] One recently observed possibility is the use of names (Pardo et al., 2024).

It is now time to look more closely at what incredulous judges of Turing Tests say that failing computers are missing.

6 Requirements for Passing

One of the most all-compassing characterizations of Intelligence is that of McCarthy and Hayes (1969, pp. 465–466):

> . . . an entity is intelligent if it has an adequate model of the world (including the intellectual world of mathematics, understanding of its own goals and other mental processes), if it is clever enough to answer a wide variety of questions on the basis of this model, if it can get additional information from the external world when required, and can perform such tasks in the external world as its goals demand and its physical abilities permit.

It is not at all clear that LLMs satisfy any aspect of this definition (Hutson, 2023). At best, LLMs seem to be "clever enough to answer a wide variety of questions", but not on the basis of any "model of the world".

Moreover, Turing Tests are vague. There is no detailed specification for how to conduct or judge one. Non-expert judges might think that LLMs pass, uncritically taking the LLM-generated output at face value. But they shouldn't, not least because of hallucinations and the lack of care about truth. Expert judges might be impressed by LLMs' linguistic abilities, but still be skeptical. Such judges will want the tests to be more stringent; they will want to test for other features of Intelligence that LLMs don't seem to have (yet).

What else is needed for cognition? What must be added to an LLM's capabilities to make it a full natural-language *understanding* program as well as a computational theory of general intelligence? What should be in the Chinese Room's rule book? Quite a few things, some of which LLMs already do, but many that they don't.[99]

1. An incredulous Turing Test-passing AI system must take discourse as input, not isolated sentences. LLMs clearly satisfy this. What is not so clear is whether they can deal with the compositional structure of individual sentences: LLMs' token-prediction seems to be primarily linear, not structured (Marcus and Davis, 2019, p. 87).

[99]The following list is based on, and extends, the list in Rapaport 1995, §2.1.2, pp. 50–51. (Cf. Rapaport 2023, §18.10.) It is not intended to be exhaustive; for related lists, see Marcus and Davis 2019; Marcus 2020, 2024l,k.

2. It must understand ungrammatical input and recover from misinterpretations. LLMs seem to be able to do this, although not always satisfactorily.[100]

3. It must be able to *understand* plans, especially the speech-act plans of users. LLMs seem to be able to do this, though again not always satisfactorily.

4. It must be able to *make* plans (Michie 1971, pp. 101–103; Kambhampati 2023) and to think ahead (Newport, 2024), both in general and for natural-language *generation*, in particular in order to ask and answer questions and to initiate or end conversations. To do this, it must have intentions (Bender and Koller's "purpose" or "communicative intent"; recall §3.8.1, above). One aspect of the Eliza Effect is that the audience assumes that the producer has such a purpose or intent. But LLMs don't. There is no evidence that LLMs can make plans; they are reactive only. More generally, an AI system must be an agent that can act in the real world. In particular, it must be an agent that can act (or even communicate) without being prompted to (or asked a question); i.e., it must have initiative and motivations.[101] Moreover, even if LLMs

    > can consciously decide on a plan, and even if we agree they can then devise a plan, these LLMs must be able to determine that it is moving forward in its plan and that means it must be able to determine what is now (or what has become) true based on the actions it is taking. (Saba, 2023)

   This is also something that they cannot (yet) do. Even more generally, an AI

[100]For example:

> I asked [ChatGPT] to find anagrams for the word 'threads.' It came up with nonsense words that didn't even have the same letters as the original. And it got "defensive" when I pointed out its mistakes. [It] list[ed] the anagrams I found after it told me it couldn't find any anagrams. It then told me I was correct and proceeded to define the words I had listed. Given sufficient memory for a database of English words, in 1980 I could have programmed my pre-IBM CPM desktop to correctly find the anagrams. . . . "Using all seven letters only once", it comes up with 'struther', which has eight letters, is not a word, contains no a and no d, and, when I point out that there is no u in 'threads', in perfect English apologizes and says, OK, you're correct, in that case the anagram is—struther. It didn't even learn to eliminate its last incorrect response. It took self control for me not to type in "are you on drugs?" And yet, if you tell it to write a six page paper explaining the origins of the Russian revolution, it's difficult to tell that a human didn't write it. (Michael Seymour, personal email, 2 January 2024)

For similar examples, see von Hippel 2023; Marcus 2024f.

[101]Thanks to Johan Lammens for emphasizing this to me. As Gleick (2024, p. 30) says, "Agency is what distinguishes us from machines."

system must be capable of decision-making (cf. Luccioni and Marcus 2023). And all of this requires being goal-oriented (cf. Jordan 2019).

5. Part of decision making is the ability to make inferences and revise beliefs (Kambhampati, 2023; Mitchell et al., 2023):

> There is no principled solution to hallucinations in systems that traffic only in the statistics of language without explicit representation of facts and explicit tools to reason over those facts. (Marcus, 2024c)

   Although LLMs seem to be able to make inferences, they do not do so on the basis of a reasoning module, and so they can just as easily "infer" a falsehood as a truth (Berglund et al. 2024; Mirzadeh et al. 2024). And although they can seem to revise their beliefs (when they say things like, "I misunderstood you", e.g., Włodarczyk 2023), they do not do so on the basis of a belief revision module.[102] Belief revision (as well as other aspects of understanding) requires an awareness of what the system is saying (and why it says it).

6. An AI system must be able to remember what it is told, what it has learned, and what it used to believe in cases where it has changed its beliefs. LLMs lack anything that we could call a belief (for arguments to that effect, see Levinstein 2023b; Schulte 2023, p. 46), and it is not at all clear that current LLMs can remember what they are told (but see Metz 2024b; Roose 2024b).

7. Related to this is the need for the system to be aware of what it is saying or doing, a sort of "higher-order consciousness" or feedback from its own output. The "psychological" level (see §4.3, above) must influence its own processing at the Design or Physical Level (even if that processing is mere donkey work), either by conscious awareness or by unconscious influence. Instinct and unconscious experience can give us information about the world we live in, but *thinking about* the instinct or experience—making it conscious—allows us to talk and theorize about it, and to consciously *use* that information (Rapaport, 2024b, §2.5, p. 21). This is needed for Intelligence; LLMs lack it.

8. To have and use a reasoning module, an AI system must care about truth. An AI system needs (a) the ability (and the desire) to justify what it says,[103] (b) knowledge of the difference between fiction and non-fiction, and (c) the

[102] E.g., Martins and Shapiro 1988.

[103] "If a mental state 'says' that things are thus-and-so, there should be something that justifies it in 'saying' that this is how things are. Otherwise what the mental state is 'saying' would be entirely arbitrary—not the kind of thing that could ground *justified* beliefs" (Kriegel, 2024, p. 466).

intention to write or speak truthfully or accurately (as well as being honest about when it is intentionally writing or speaking fictionally). LLMs'

> deepest flaw is the absence of the most critical capacity of any intelligence: to say not only what is the case, what was the case and what will be the case—that's description and prediction [which is "The crux of machine learning"]—but also what is not the case and what could and could not be the case. Those are the ingredients of explanation, the mark of true intelligence [that "human-style thought is based on"]. (Chomsky et al., 2023)

9. Language learning, linguistic negotiation (Rapaport, 2003a), and the ability to have a real conversation require an AI system to be able to interact with its interlocutors, to exhibit joint attention, and to construct a model of the user's beliefs (Bender and Koller 2020, p. 5190; Bisk et al. 2020, p. 8722, col. 2). There is no evidence that LLMs can do any of these. Related to this, an AI system must be capable of *experiential* learning: Are there "things that *cannot* be learned (about language) by merely reading large bodies of text data?" (Sahlgren and Carlsson 2021, §2.3; again, cf. Jackson 1986). After all, compare learning French in artificial situations in school with using it in real-life situations in France.

   > Language learning continues for a speaker's whole lifetime: the grammar of human languages evolves, and humans flexibly adapt to novel linguistic situations … (Bommasani et al., 2022, p. 26)

   > Consciousness depends on a brain's ability to maintain a constantly updated conception of itself as a distinct entity interacting with a model of the external world. The layers of neural networks that make up systems like ChatGPT, however, are static: once they're trained, they never change. ChatGPT maintains no persistent state, no model of its surroundings that it modifies with new information, no memory of past conversations. (Newport, 2023)

   Note, however, that Bubeck et al. (2023, §§5.2.1, 5.2.2) suggest that GPT-4 is capable of learning from current interactions.

10. It must have background knowledge, including "world knowledge" and "commonsense knowledge". By this, I mean that it must have a knowledge base of true statements ("beliefs") about the world and about language, not just the ability to statistically predict words that form sentences that look to us

as if they constitute such knowledge. Although LLMs appear to have some world or background knowledge (though not common sense), they only have what they can probabilistically predict from their training. They do not have an explicit knowledge representation system. Such knowledge should not merely be a list of unrelated propositions. The knowledge must be organized as a *model* of (relevant aspects of) the world (Thorpe, 1989). Even if LLMs automatically "induce" models in some fashion, it's not clear that they *use* those models.

11. An AI system must also be able to learn about the world and about language. Current LLMs "learn" statistical correlations, but cannot "learn" new things (OpenAI, 2023, p. 18).[104] To the extent that LLMs are taken to be Intelligent, the way that they become so differs considerably from the way that humans do. How does a *human* become Intelligent? Turing (1950, §7, p. 455) suggests "three components":

    > (a) The initial state of the mind, say at birth,
    > (b) The education to which it has been subjected,
    > (c) Other experience, not to be described as education, to which it has been subjected.

    And he envisages a Test-passing program as including "two parts[:] the child-programme and the education process" (p. 456). If these are (a) and (b), what about (c)? Let's consider what might correspond to these in the case of an LLM (or a successor):

    (a) We might understand the "initial state of the mind" to include the LLM's initial training, but that training is better understood as falling under the "education process". It is more plausible to understand a human's initial state in the sense of what is traditionally called "innate knowledge". There is a long tradition of this in philosophy and cognitive science, of course, from Plato to Kant to Chomsky. Stanislas Dehaene (2020, esp. Ch. 5) has argued forcefully for its importance in the case of humans. His list includes what he calls the object concept, the number sense, the intuition of probabilities, knowledge of animals and people, face perception, and the language instinct. (See also Marcus and Davis 2019, pp. 25–26; Levy 2023.) These might be considered as, roughly, the "base cases" of a "recursive" process of learning. I think that it's fair to say that this is missing from LLMs. At best, what might count as the

[104] Bringsjord et al. (2018) argue that machine-learning systems don't "learn" anything at all. See also Rothman 2024.

initial state of the "mind" for an LLM are its statistical, deep-learning algorithms that await textual input to process.[105]

(b) LLMs have had at least two kinds of "education": There is, first, the statistical modeling that it undergoes based on an immensely large corpora of text. Second, there is some explicit teaching stemming from human training (Metz, 2023a; Lu, 2024). However, there is no explicit "teaching" of specific subject matter, such as, for example, mathematics (Marcus, 2023b). (Note that LLMs often make simple arithmetic mistakes when they shouldn't—i.e., when they are not trying to fool a Turing Test judge.)[106] Both (a) and (b) may require a GOFAI knowledge-representation and reasoning system.

(c) Turing's "other experience" presumably includes what might be called "real life" experience. This is certainly something not explicitly part of current LLMs whose training stops on a specific date and is not updated thereafter. And LLMs are not active or independent participants in daily life. Component (c) probably requires the system to be a multimodal, active, embodied agent.[107]

12. To have a model, a cognitive entity must have internal representatives of external objects as part of its knowledge representation system. To have such avatars, it must be able to perceive (it must be multimodal).[108] Perhaps it must be embodied, although Bisk et al. (2020, p. 8722, col. 2) suggest that "virtual worlds" (i.e., internal, simulated environments) might suffice. As noted with the whales (recall §5.3), it also needs to be embedded in a social context (cf. Humphrey 1976; Bisk et al. 2020). LLMs cannot do this (yet):

> Text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. It can't have been, because the training data never included sharing thoughts with a listener, nor does the machine have the ability to do that. (Bender et al., 2021, p. 616, col. 2)

---

[105] For a contrary point of view and an enlightening discussion of innateness in the context of deep learning, see Buckner 2024.

[106] For recent experiments along these lines, see Metz 2024a. LLMs currently don't seem to do so well at math (Lohr, 2024). If they need to do better, they may well need some help from GOFAI techniques, or at least access to a calculator.

[107] For more on the education of AI systems, see Coffey 2024; Whang 2024.

[108] Note that Helen Keller, albeit deaf and blind, could still perceive (Keller, 1905; Rapaport, 2006, 2011).

Bender et al. make two important points. The first is the statement of what LLMs lack. The second is an interesting explanation of that lack, interesting because it suggests that if an LLM could be trained with shared thoughts with an interlocutor, it could have the things that it now lacks. This is essentially part of the learning that an LLM "child" would have.

13. An AI system needs to understand causality and uncertainty (cf. Jordan 2019).

14. And, of great importance, any biases that such systems show are embedded in the data on which they are trained (OpenAI, 2023, p. 7). Intelligence requires the computer to be "aware" of such biases and to be able to overcome them without external prompts. And the difficulty of doing this is one of the chief dangers of (current) LLMs.

So I think we can be confident that, despite seeming to pass a Turing Test—or despite passing a credulous Turing Test—LLMs lack many of the features required for Intelligence (or for passing an incredulous Turing Test).

Peter J. Denning believes that many of these things are unattainable:

> Have we become so mesmerized by LLMs we do not see the rest of what we do in language? … We build relationships. We take care of each other. We recognize and navigate our moods. We build and exercise power. We make commitments and follow through with them. We build organizations and societies. We create traditions and histories. We take responsibility for actions. We build trust. We cultivate wisdom. We love. We imagine what has never been imagined before. We smell the flowers and celebrate with our loved ones. None of these is statistical. (Denning, 2023b, p. 27, col. 1)

Of course, it's not clear that these are necessary conditions for Intelligence. But note that even if "None of these is statistical", some or all of them might yet be *computable*. Yet Denning thinks that they are not even that:

> An analogy familiar to computer scientists is the gap between Turing machine-computable functions and all functions: the machines are a countable infinity, the functions are an uncountable infinity. There are not enough LLMs to handle all the functions visible in human interactions. (Denning 2023b, p. 27, col. 2; cf. Denning and Lewis 2019)

But, in the absence of a mathematical proof of non-computability, whether this is the case is an empirical question (Rapaport, 2024b).

LLMs arguably pass Turing Tests when their users accept the output as true, leading to Turing's "words will alter" state of affairs. But before "general educated opinion" alters, LLMs will need to do a lot more; the current bar for passing is too low. There is more to Intelligence than stochastic parroting (even if what's needed is still donkey work).

7 Lessons to Be Learned from LLMs

7.1 LLMs and AI Research

How do LLMs fit into Shapiro's three AI research categories? One of the most fascinating things about these systems is the quantity and kinds of cognitive activity that turn out to be computable in this statistical version of the engineering approach:

> ... what we should conclude is that tasks—like writing essays—that we humans could do, but we didn't think computers could do, are actually in some sense computationally easier than we thought. (Wolfram, 2023, p. 39)

Recall Turing's nicely neutral phrase 'intellectual capacities' (see §3, above). What LLMs seem to show is that some such capacities are "buried" in our vast textual database. This is consistent with locating the Intelligence of the Chinese Room Argument at least partially in the rule book. But only *some* intellectual capacities are so buried; LLMs by themselves are not yet capable of *all* cognition. They will probably need to be supplemented by GOFAI techniques (perhaps among other things).[109] There is a similarity between, on the one hand, LLM processing and Daniel Kahneman's (2011) notion of "fast thinking", and, on the other hand, GOFAI processing and Kahneman's notion of "slow thinking", as well as the need for both (Bubeck et al., 2023, pp. 91, 94). As Roger Schank showed many years ago (using GOFAI techniques), it is possible to write a story or text (or to continue one from a prompt) by using defeasible rules about what is likely to have happened, without having to know (or "know") any of the details. But that is using a generalized form of knowledge (or "knowledge") about stereotypical situations. LLMs use an entirely different method. Floridi and Chiriatti (2020, pp. 685–686) gave GPT-3 the first sentence of a passage about an accident from a(n unfinished) novel by Jane Austen. GPT-3 continued the story, not as Austen did, but in a reasonable way. They comment that this is "Because if all you know is the occurrence and nature of the accident, it makes a lot of sense to assume that the passengers might

[109] See the references in footnote 84, above. For a possible way to extend LLMs short of GOFAI, see Lammens 2024.

have been injured" (ibid.) But GPT-3 did not use a defeasible rule to that effect; it used statistical likeliness: similar end, very different means.

Thus, Wolfram's claim (2023, p. 39), along the lines of the computational psychology approach, that LLMs' success (see footnote 14, above) "takes us closer to 'having a theory' of how we humans manage to do things like writing essays, or in general deal with language" doesn't follow. Yes, it *might* be the case that much of human essay writing, to take his example, is accomplished by our brain's neural network doing statistical processing, but that's at best an empirical hypothesis. It is possible to conclude from a study of how *humans* accomplish some cognitive task that a *computer* might be programmed to do it in that way (as in Newell, Shaw, and Simon's 1958 work on the Logic Theorist). But we should be cautious about concluding from a study of how an LLM writes essays, for example, that that's how *humans* do it.

When *typical* users assign a task to an LLM, they *expect* it to "think about" that task and to come up with a "reasoned" response. They don't *expect* it to merely generate a grammatical set of words that are statistically related to the task statement. Yet that's what LLMs do. The point here is that, for current LLMs, that's *all* that they do (at least, *qua* LLMs; LLMs can be embedded in a system that *can* do other things; cf. Shanahan 2024). Descartes observed

> that although machines can perform certain things as well as or perhaps better than any of us can do, they infallibly fall short in others, by the which means we may discover that *they did not act from knowledge*, but only from the disposition of their organs. (Descartes, 1637, Part V, p. 116, my italics)[110]

Although Descartes seems here to be worried about the *physical* actions of a machine, we can still ask if an LLM is "acting from knowledge". It *is* acting from statistical analysis of its training data; is that "knowledge"? The open issue is whether such statistical processing is a way of coming up with a "reasoned" response. Does an LLM "find" a reasoned response in its statistical analysis of the texts that it was trained on?

One aspect of our paradox concerns the nature of the algorithms at their lowest level: the level of the workers whom Adam Smith (1776) said were as "stupid and ignorant as it is possible for a human creature to become",[111] the level of "drudgery" that Babbage "deplored" (Stein, 1984, pp. 51–52), the level that Turing called "donkey work", the "mechanistic" level described by Dennett (1975, p. 179) as "where the homunculi … need only the intelligence to pick the larger

[110] Interestingly, one of the things that Descartes thought that machines fell short in was "arrang[ing] its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do." LLMs do not fall short here!

[111] https://tinyurl.com/smith1776ch01c-2

of two numbers when directed to". Should those algorithms be written so that they capture what Intelligence or natural-language understanding "really" is? Or is it sufficient for them to produce output that a judge says is Intelligent? The latter would be AI as engineering; the former would be AI as computational psychology. The paradoxical responses to LLMs suggest that, despite merely getting the job of language processing done by whatever means work, there is still a desire for a scientific understanding of the cognition involved in language understanding.

7.2 The Nature of Intelligence

One lesson we seem to be learning from LLMs concerns our views about the nature of Intelligence. Can the judge in a Turing Test be wrong about a computer's cognitive abilities? Although Turing's answer is traditionally taken to be "no", and Searle's is "yes", a more cautious interpretation is consistent with the response-dependent interpretation, and certainly consistent with Turing's views about "the use of words and general educated opinion": Whether the judges are wrong depends on what *they* take Intelligence to be (and, in the case of LLMs, we its users are the judges). The Turing Test does not answer the ontological question of whether a computer *is* Intelligent or what Intelligence *is*. At best, it shows what *we* (as the judge in a Turing Test) *think* that it is; it is epistemic in nature.

One of the standard problems of AI, as Bertram Raphael is famously said to have observed, is that "AI is a collective name for problems which we do not yet know how to solve properly by computer",[112] from which it follows that once we do know how to solve them, they are no longer AI! Once we know how something works, we see it as donkey work, not Intelligence. Raphael is not alone in this:

> ... two decades ago, if we'd seen a system behaving as LLMs do *without knowing how it worked*, we'd have taken this behavior as fairly strong evidence for intelligence and consciousness. (Chalmers, 2023a, my italics)

Chalmers seems to be saying that *because* we know how it works, we know that it's *not* intelligent. He does not believe that LLMs have passed the Turing Test, although "they're not so far away. Their performance often seems on a par at least with that of a sophisticated child. And these systems are developing fast" (ibid.). But sophisticated children *are* Intelligent and capable of thinking, so haven't LLMs that are "on a par" with them passed a Turing Test?

Bennie Mols (2023) quotes Melanie Mitchell as saying almost the same thing (my italics):

[112] Cited in Michie 1971, p.101

> Before Deep Blue beat Kasparov, people honestly believed that playing chess requires general intelligence; *now we know that it does not.* Large language models demonstrate that language understanding is not needed to generate humanlike text. ... [B]uilding AI systems refines our understanding of what intelligence is.

That last sentence is ambiguous. It could mean that Intelligence is *different from* what LLMs do. But it could mean that Intelligence turns out to *be* what LLMs do! If there is a difference between (mere) natural-language *processing* and natural-language *understanding*, then LLMs do show that natural-language processing without natural-language understanding can "generate humanlike text". The question remains of what the difference between them is (as Mitchell notes).

Is it fair to say that natural-language understanding or playing chess don't *require* Intelligence?

> While a human player uses memory, experience, high-level abstract reasoning, pattern recognition, and intuition to cast his or her mind over the board, a chess engine does not really understand the game at all, it simply uses its power to calculate and then makes a decision following a complex set of hand-crafted rules laid down by its programmers. (Labatut, 2023, pp. 298–299)

There does seem to be a way to play chess (the neural-network machine-learning way) that one could argue does not require the *kind* of Intelligence that *humans* apply to playing the game. But do we really *use* memory, etc., to play chess? Or do we only *think* that we do? (Where what we "really" do is fire neurons.) Why is "simply" calculating *not* understanding? Who is to say that what *we* think of as the human, non-statistical, non-machine-learning way of playing isn't, at bottom, the same as the computer's way? After all, at bottom, the human way is also based on a neural network. Perhaps, then, the machine-learning way *is* Intelligent. Suppose that the computer "simply" calculates, but is also able to Intelligently discuss the game. Would that not show that it understands the game (even if it can't explain how or why it makes certain moves)? (This, of course, might only be the case for a neural-network machine-learning computer, not for a GOFAI computer with a cognitively based algorithm.) This is the donkey-work-as-red-herring point (§3.8.4): The *manner* of playing (memory, etc., on the one hand, vs. calculating or statistical machine-learning, on the other) does not really matter. Yet it *may* matter if purely donkey-work/statistical algorithms don't suffice for the needs listed in §6.

The same goes for language. We might think that an LLM isn't processing (or understanding) language the way we do. But perhaps we should say that it isn't processing (or understanding) the way that we *think* we do. In any discussion of these alternatives, we have to keep in mind one striking fact: In the case of the gold standard of (human) language understanding, it is all accomplished by neuron firings, which really don't seem very different from donkey work:

> Engineers who know the mechanism of advanced robots most intimately will be the last to admit they have real minds. From the inside robots will indisputably be machines, acting according to mechanical principles, however elaborately layered. Only on the outside, where they can be appreciated as a whole, will the impression of intelligence emerge. A human brain, too, does not exhibit the intelligence under a neurobiologist's microscope that it does participating in a lively conversation. (Moravec, 1998, p. 10)

Again, recall Leibniz's mill.

Bridewell and Isaac (2021) have elevated a version of Raphael's observation to a research strategy: the "apophatic" method of characterizing a phenomenon by saying what it is *not*. If a computational model of some feature of cognition (in their case, that feature is consciousness) succeeds, that shows that we have *not* successfully understood it: "modeling success demonstrates what consciousness is not" (Bridewell and Isaac, 2023):

> If natural language production and understanding in human beings is not algorithmic, one of the great merits of algorithmic attempts to simulate these processes will lie precisely in their failure to capture non-algorithmic principles with algorithmic methods. (Searle, 1986).

But that LLMs don't understand *because* all they do is predict the next term is *not* a good argument: For one thing, it would follow that *we* don't understand, because all *we* do is fire neurons. For another, the absence of other features of understanding (see §6) suggests that *only* predicting the next term is *not* sufficient for understanding. The apophatic methodology *assumes* that computation is *insufficient* for understanding, because if computation yields understanding, then it is not understanding!

7.3 Risks

Another lesson to be learned from LLMs (or reminded of by them) is that "devices that use heuristics to create the *illusion* of intelligence present a risk we should not accept" (Parnas, 2017, p. 5, my italics). Although there are many such risks, one of the biggest is the bias problem arising from how these systems are trained. Granted, the training data is huge. But …

> ... Despite claims by large-language-model enthusiasts that their training sets are all-encompassing, the conversations embodied into the neural network come from a particular crowd. The crowd does not encompass all humanity. ... Chatbot models are notoriously biased toward the conversations among the well-educated and well-off even within rich countries. (Denning, 2023a, p. 27)

Any trust we might put into such an AI must be tempered by an awareness of the possibility of such bias. Of course, the same must be said for humans: Any trust we put into what we read or hear someone say must be similarly tempered. We should always be at least slightly skeptical (Rapaport, 2023, §2.4.4). And we need to be *able* to think critically—and to *actually* think critically!—in order to overcome paralyzing skepticism (Graham and Metaxas, 2003; Singer, 2023; Waxman, 2024).[113]

As many stories about the wishes granted by genies make clear, just as you should be careful about what you wish for, you should also be careful about what you ask an LLM. Whenever you deal with one, you are in the position of a Turing Test judge. But if you're not trying to find an LLM's weakness that reveals itself as a mere computer, then you need to ask it questions in a way that will raise the probability of getting a reasonable answer. (There is now a small industry offering such advice; e.g., Chen 2023.)

8 Summary

> When the moral courage to decide and differentiate between fraud and reality begins to melt away, that marks the end of life itself, of formed opinions, of values, of any improving deed, and the corruptive process of moral skepticism begins its awful work.
> —Thomas Mann, *The Magic Mountain* (1924, p. 657)

8.1 Machines and Thinking

The usual informal interpretation of the Turing Test is that it is a measure or sign of thinking or Intelligence (Claim 1). But that's not what Turing said. It is, of

[113] William Perry's (1970; 1981) "Scheme of Intellectual and Ethical Development" can shed some light on these issues. On his scheme, "Dualists" believe things uncritically, before becoming "Multiplists", who believe that all opinions are equally good. Both of these positions seem to describe many non-expert users of LLMs. It is "Contextual Relativists"—who try to understand things relative to their contexts—who have a chance to become expert users. For discussion of Perry, see Rapaport 2018; Rapaport 2023, §2.6, pp. 22–23; and https://tinyurl.com/phics-perry.

course, extremely tempting to infer from passing a Turing Test that the entity "is" Intelligent (to make an ontological inference from an epistemic premise). But that's not an inference that Turing drew.

That machines might think was an oxymoron at the time when Turing developed his "test". To show that it wasn't, he devised his Imitation Game thought experiment (see Gonçalves 2023). More precisely, he believed that the (linguistic) output of some machines (digital computers, in particular) could be indistinguishable from the (linguistic) output of humans, i.e., what human interlocutors call 'thinking':

> He never really said that machines could be said to be thinking if they played the Imitation game successfully. He merely said that successful imitation was the real issue. (Schank, nd, p. 2)

He predicted that successful imitation would be *taken as* Intelligence (our use of words would alter), and also that our notion of Intelligence would change to include such imitation (our educated opinion would alter). "Thinking machine" would no longer be an oxymoron.

The gold-standard method for producing Intelligence is the human brain. (If you widen the notion of Intelligence to other animals, then other biological brains should be included. If there are extra-terrestrials, then their exobiological methods could be added to this list.) AI has offered computational methods such as symbolic GOFAI algorithms, LLM algorithms, and perhaps hybrid GOFAI/LLM systems.

Suppose that such different varieties of internal processing all produce indistinguishable output. It is neither the processing nor the output by themselves that determine an entity's Intelligence. It is our response to output that matters, and Turing Tests are that response. Both internal processing and external output *are* important, of course: The external output must be such that we judge it to be Intelligent, and the internal processing must be such that it produces that output. Different judges will make different decisions (or if 'decision' is too intentional a term, then let us say that different judges will respond differently to the output). Credulous judges will be (perhaps overly) lenient in their judgments; incredulous judges will be (perhaps overly) skeptical. But it doesn't matter *how* the output is produced (as long as the processing is computable, if it's to be considered AI). Donkey work *might* be capable of producing output judged to be Intelligent. If so, we will just get confused if we see ("unIntelligent") donkey work producing output judged to be Intelligent: "you are not allowed to go behind the scenes and criticise the method [the internal processing], but must abide by the scoring on correct answers [the external output]" (Max Newman, in Turing et al. 1952, p. 496). Yet it's possible that the internal processing might matter, that there must be some structures in the internal processing that mirror (and not merely result in) features such as those listed in §6.[114]

In light of all this, it is time to resolve the inconsistency in our initial triad:

Claim 1 should be revised to say merely that if an entity is judged (or taken) to have passed a Turing Test, then the judge is willing to treat that entity as Intelligent.

Claim 2 should be qualified to say that some LLMs are sometimes judged to have passed a Turing Test. A given LLM might pass some Turing Tests and not others, if the judges differ in their responses to it. But given the general untrustworthiness of much current LLM output, judges should be urged to err on the side of incredulity. The same, of course, holds true for our responses to our fellow humans.

Claim 3: Claim 3a (that current LLMs are stochastic parrots) is probably true. But, in view of the revised Claim 1, the truth value of Claim 3b (that stochastic parrots are not Intelligent) depends on the judge. The method of stochastic parroting by itself does not rule out its output being judged Intelligent.

This triad is not inconsistent.

It is a fact that (some) LLMs pass (some) credulously judged Turing Tests. Yet I still don't want to say that they are Intelligent. And this is because I think that tests for Intelligence have to include tests for the features in §6. And that, in turn, might well mean that those features must be explicitly in the internal processing (perhaps in a GOFAI way). To the extent that they are not produced by LLM algorithms, LLMs won't pass incredulously judged Turing Tests. So Turing may have been wrong if he meant that there really is nothing more to Intelligence than what a judge finds in the output. (Of course, he may not have meant that.) But he was sadly right that our words would alter, that people would take the output as signs of Intelligence.

8.2 The Real Dangers of LLMs (and AI)

Many commentators have pointed to the alleged dangers of artificial general intelligence and the Singularity. But there is a more serious—because more pressing—problem: our current willingness to credulously accept Turing Test-passing entities as Intelligent. Despite their hallucinations and confabulations, "millions of people

[114]The foucs on output is consistent with the possibility of philosophical zombies. But if there must be something in the processing that "mirrors" the output, then such a system would not be a philosophical zombie.

do trust A.I. models, and their outputs are being given prominent real estate on . . . Google, . . . Facebook . . . , even in basic Microsoft Office applications" (Roose, 2024a).[115]

Turing was prescient: Whether or not LLMs get closer to, or achieve, artificial general intelligence, they are already accepted by many people as being Intelligent. And as they get better, more people will accept them, flaws and all. Indeed, they might not only be accepting of them, they might "become unable or unwilling to distinguish artificial systems from human systems" (Schwitzgebel et al., 2023, 2). And therein lies the danger, because such entities *don't* pass "aggressive" Turing Tests, yet we are all too willing to see Intelligence in their behavior. We come to trust and rely on them, when we should be more critical and less accepting of them.

The source of this danger—and it is a danger, especially given that most interactions with them will be credulous ones—has three interacting sources: (1) the fact that "The basic tool for the manipulation of reality is the manipulation of words" (Dick, 1978),[116] (2) LLMs' ignorance of truth, and (3) the combined interaction of the Eliza Effect, response dependence, and the Intentional Stance. Dennett, too, was prescient: In 1985, he wrote:

> The problem of overestimation of cognitive prowess, of comprehension, of intelligence, is not, then, just a philosophical problem, but a real social problem, and we should alert ourselves to it, and take steps to avert it. (Dennett, 1985, p. 140)

Sadly, we did not. We underestimated how soon it would happen.

It may or may not be the case that LLMs as currently implemented will achieve artificial general intelligence. Perhaps it will require GOFAI to do so. But suppose that, in the limiting case, an LLM-based system does so. We would still need to treat its output critically and not with blind acceptance of its Intelligence, in precisely the same way that we should treat the output of other humans. All the harms of LLM-based AIs are also possible harms of *human* linguistic interaction, but humans can override them (even if they don't). Similarly, even if LLMs learn

[115]Google now gives an LLM response as the first result when searching. I recently did a Google search for the actor David Alan Grier. Google's top reply was an "overview" that told me that "David Alan Grier is an American actor, comedian, and associate professor of science and technology policy at George Washington University". I was rather impressed until I checked further and discovered that there are *two* David Alan Griers: One is an actor; the other is a professor. Only in small print at the end of the "overview" was there a caveat: "Generative AI is experimental".

[116]Cf. Hannah Arendt (1974, cited in Berkowitz 2024): "If everybody always lies to you, the consequence is not that you believe the lies, but rather that nobody believes anything any longer. . . . And a people that no longer can believe anything cannot make up its mind. It is deprived not only of its capacity to act but also of its capacity to think and to judge. And with such a people you can then do what you please."

from uncurated data "exactly" as humans do, LLMs can, theoretically, be trained on (positively) curated data: Both LLMs and humans need to be better educated. "You've got to be carefully taught".[117]

> Goat: "Looks like artificial intelligence is really getting better."
> Rat: "What can we do about regular intelligence?"
> Goat: "I think that's stuck where it is."
> Rat: "Scientists need to prioritize."
> —"Pearls before Swine" comic strip, 17 June 2023[118]

References

Andrews, K. and B. Huss (2014). Anthropomorphism, anthropectomy, and the null hypothesis. *Biology and Philosophy 29*, 711–729. Page refs. to https://tinyurl.com/andrews-huss2014.

Argamon, S., M. Koppel, et al. (2003). Gender, genre, and writing style in formal written texts. *Text & Talk 23*(3), 321–346.

Astobiza, A. M. (2024). Do people believe that machines have minds and free will? Empirical evidence on mind perception and autonomy in machines. *AI and Ethics 4*, 1175–1183.

Banfield, A. (1982). *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Boston: Routledge & Kegan Paul.

Barr, A. (1983). Artificial intelligence: Cognition as computation. In F. Machlup and U. Mansfield (Eds.), *The Study of Information*. New York: Wiley. Page ref. to https://tinyurl.com/barr1983.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bender, E. M. and A. Koller (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proc. 58th Annual Meeting Assn. Comp. Ling.*, pp. 5185–5198.

Berglund, L., M. Tong, et al. (2024, 26 May). The reversal curse: LLMs trained on 'A is B' fail to learn 'B is A'. https://arxiv.org/abs/2309.12288.

Berkowitz, R. (2024, 4 August). On fake Hannah Arendt quotations. *Amor Mundi*. https://tinyurl.com/berkowitz2024.

Bisk, Y., A. Holtzman, et al. (2020, 16–20 Nov.). Experience grounds language. *Proc. 2020 Conf. Empirical Methods Nat. Lang. Processing*, 8718–8735. https://tinyurl.com/bisketal2020.

Bjarnason, B. (2023, 4 July). The LLMentalist effect: How chat-based large language models replicate the mechanisms of a psychic's con. https://softwarecrisis.dev/letters/llmentalist/.

[117] https://tinyurl.com/2p95zp56
[118] https://www.gocomics.com/pearlsbeforeswine/2023/06/17

Boden, M. A. (2018). *AI: A Very Short Introduction*. Oxford University Press.

Bommasani, R., D. A. Hudson, et al. (2022). On the opportunities and risks of foundation models. https://arxiv.org/abs/2108.07258v3.

Borges, J. L. (1941). The library of Babel. In A. Kerrigan (Ed.), *Ficciones*, pp. 79–88. New York: Grove Press, 1962.

Boswell, J. (1791). *The Life of Samuel Johnson*. London: Henry Baldwin. https://tinyurl.com/boswell1791.

Brachman, R. J. and H. J. Levesque (2022). *Machines Like Us: Toward AI with Common Sense*. Cambridge, MA: MIT Press.

Bridewell, W. and A. M. Isaac (2021). Apophatic science: How computational modeling can explain consciousness. *Neuroscience of Consciousness 7*(1).

Bridewell, W. and A. M. Isaac (2023, 20 Oct.). Computational modeling and consciousness 5: Response to comments. *The Brains Blog*, https://tinyurl.com/bridewell--isaac20231020.

Bringsjord, S., N. S. Govindarajulu, et al. (2018). Do machine-learning machines learn? In V. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2017*, pp. 136–157. Cham, Switzerland: Springer.

Broad, W. J. (2024, 23 Dec.). How hallucinatory A.I. helps science dream up big breakthroughs. *New York Times*. https://tinyurl.com/broad2024.

Bubeck, S., V. Chandrasekaran, et al. (2023, 13 April). Sparks of artificial general intelligence: Early experiments with GPT-4. https://arxiv.org/abs/2303.12712.

Buckner, C. J. (2024). *From Deep Learning to Rational Machines*. Oxford University Press.

Cappelen, H. and J. Deaver (2021). *Making AI Intelligible*. Oxford University Press.

Carroll, L. (1871). *Through the Looking-Glass*. http://www.gutenberg.org/files/12/12-h/12-h.htm.

Carter, D. (2023, 13 May). There are no 'emergent abilities' in LLMs. *Better Programming*. http://tinyurl.com/carter20230513.

Castañeda, H.-N. (1966). 'He': A study in the logic of self-consciousness. *Ratio 8*, 130–157.

Castañeda, H.-N. (1967). Indicators and quasi-indicators. *Amer. Phil. Qtly 4*, 85–100.

Chakraborty, T. and S. Masud (2024, Online 29 August). The Promethean dilemma of AI at the intersection of hallucination and creativity. *Communications of the ACM*.

Chalmers, D. J. (2020, 30 July). GPT-3 and general intelligence. *Daily Nous*. In "Philosophers on GPT-3", A. Zimmermann (ed.), https://dailynous.com/2020/07/30/philosophers-gpt-3/.

Chalmers, D. J. (2023a, 9 August). Could a large language model be conscious? *Boston Review*. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/.

Chalmers, D. J. (2023b). Does thought require sensory grounding? From pure thinkers to large language models. *Proc. & Addr. Amer. Phil. Assn 97*, 22–45.

Chen, B. X. (2023, 25 May). Get the best from ChatGPT with these golden prompts. *New York Times*.

Chesterton, G. (1907, 2 January). Books of the day: An edition of Dumas. *The [London] Daily News*. https://quoteinvestigator.com/2023/08/21/best-writers/.

Chiang, T. (2002). Story of your life. In *Stories of Your Life and Others*, pp. 91–145. New York: Vintage.

Chiang, T. (2019). The lifecycle of software objects. In *Exhalation*, pp. 62–172. New York: Alfred A. Knopf.

Chomsky, N., I. Roberts, and J. Watumull (2023, 8 Mar.). The false promise of ChatGPT. *New York Times*.

Churchland, P. M. (1984). *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind*. Cambridge, MA: MIT Press.

Coffey, L. (2024, 18 January). AIs enrolling as students in Michigan university's experiment. *Inside Higher Ed*. https://tinyurl.com/coffey2024.

Coghlan, S. (2024). Anthropomorphizing machines: Reality or popular myth? *Minds and Machines 34*(25).

Colombo, M. and G. Piccinini (2023). *The Computational Theory of Mind*. Cambridge University Press. Page ref. to private pre-print.

Copeland, B. J. (2023, Jul./Sep.). Early AI in Britain: Turing et al. *IEEE Annals of the History of Computing 45*, 19–31.

Davis, M. D. (2017). Universality is ubiquitous. In J. Floyd and A. Bokulich (Eds.), *Philosophical Explorations of the Legacy of Alan Turing*, pp. 153–158. Springer.

Dehaene, S. (2020). *How We Learn*. New York: Penguin Books.

Dennett, D. C. (1971). Intentional systems. *J. Phil. 68*, 87–106.

Dennett, D. C. (1975). Why the law of effect will not go away. *J. Theo. Soc. Behav. 5*(2), 169–188.

Dennett, D. C. (1985). Can machines think? In M. Shafto (Ed.), *How We Know*, pp. 121–145. Harper & Row.

Dennett, D. C. (2023a). *I've Been Thinking*. New York: W.W. Norton.

Dennett, D. C. (2023b, 16 May). The problem with counterfeit people. *The Atlantic*.

Denning, P. J. (2023a, June). Can generative AI bots be trusted? *Commun. ACM 66*(6), 24–27.

Denning, P. J. (2023b). The smallness of large language models. *Commun. ACM 66*(9), 24–27.

Denning, P. J. and T. G. Lewis (2019). Intelligence may not be computable. *Amer. Sci. 107*(6), 346–349.

Denning, P. J. and B. S. Rousse (2024). Can machines be in language? *Commun. ACM 67*(3), 32–35.

Descartes, R. (1637). Discourse on method. In E. S. Haldane and G. Ross (Eds.), *The Philosophical Works of Descartes*, pp. 79–130. Cambridge, UK: Cambridge Univ. Press, 1970.

Dhar, V. (2024, 21 Oct.). The paradigm shifts in artificial intelligence. *Commun. ACM, Online First*.

Dick, P. K. (1978). How to build a universe that doesn't fall apart two days later. In L. Sutin (Ed.), *The Shifting Realities of Philip K. Dick*, pp. 259–280. New York: Pantheon, 1995.

Dummett, M. A. (1976). What is a theory of meaning? (II). In G. Evans and J. McDowell (Eds.), *Truth and Meaning*, pp. 67–137. Oxford: Clarendon Press.

Ehrlich, K. (1995). Automatic vocabulary expansion through narrative context. Tech. Rep. 95-09, SUNY Buffalo Dept. of Computer Science, Buffalo, NY.

Ferrario, A., A. Facchini, and A. Termine (2024). Experts or authorities? The strange case of the presumed epistemic superiority of artificial intelligence systems. *Minds and Machines 34*(30).

Fforde, J. (2009). *Shades of Grey: The Road to High Saffron*. New York: Penguin.

Floridi, L. (2024). The ethics of artificial intelligence. *Amer. Phil. Qtly 61*(4), 301–307.

Floridi, L. and M. Chiriatti (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines 30*(4), 681–694.

Fortnow, L. (2023, 10 Apr.). Complexity and explainable AI. *Computational Complexity (blog)*. https://tinyurl.com/fortnow20230410.

Frankfurt, H. G. (2005). *On Bullshit*. Princeton, NJ: Princeton University Press.

Gabaldon, D. (1994). *Voyager*. New York: Dell.

Galbraith, M. (1995). Deictic shift theory and the poetics of involvement in narrative. In J. F. Duchan, G. A. Bruder, and L. E. Hewitt (Eds.), *Deixis in Narrative*, pp. 19–59. Hillsdale, NJ: Erlbaum.

Garnelo, M. and M. Shanahan (2019). Reconciling deep learning with symbolic artificial intelligence. *Current Opinion in Behavioral Sciences 29*, 17–23.

Garvin, P. L. (1985). The current state of language data processing. In M. C. Yovits (Ed.), *Advances in Computers, Vol. 24*, pp. 217–275. Orlando, FL: Academic Press.

Gazzaniga, M. S. (1989, 1 September). Organization of the human brain. *Science 245*(4921), 947–952.

Gleick, J. (2024). The fate of free will. *New York Review of Books 71*(1), 27–30.

Godfrey-Smith, P. (2024). Visible and invisible worlds. *New York Review of Books 71*(10), 32–34.

Goh, E. et al. (2024). Large language model influence on diagnostic reasoning. *JAMA Network Open 7*(10), e2440969.

Gonçalves, B. (2021). Machines will think: Structure and interpretation of Alan Turing's imitation game. https://tinyurl.com/goncalves2021. PhD diss., Univ. São Paulo.

Gonçalves, B. (2023). The Turing test is a thought experiment. *Minds and Machines 33*.

Gonçalves, B. (2024). *The Turing Test Argument*. New York: Routledge.

Graham, L. and P. T. Metaxas (2003). 'Of course it's true; I saw it on the internet!'. *Communications of the ACM 46*(5), 71–75.

Gubelmann, R. (2022). A loosely Wittgensteinian conception of the linguistic understanding of large language models like BERT, GPT-3, and ChatGPT. *Grazer Philosophische Studien 99*, 485–523.

Gunderson, K. (1985). *Mentality and Machines, 2nd Ed.* Univ. Minnesota Press.

Haigh, T. (2023). Conjoined twins: Artificial intelligence and the invention of computer science. *Commun. ACM 66*(6), 33–37.

Haikonon, P. O. (2020). On Artificial Intelligence and consciousness. *J. Artif. Intell. Consciousness 7*(1), 73–82.

Halpern, S. (2023, 28 Mar.). What we still don't know about how A.I. is trained. *The New Yorker*.

Han, S. J., K. J. Ransom, et al. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research 83*(101155).

Harnad, S. (1990). The symbol grounding problem. *Physica D 42*, 335–346.

Harnad, S. (1991). Other bodies, other minds. *Minds and Machines 1*(1), 43–54.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hearst, M. and H. Hirsh (2000). AI's greatest trends and controversies. *IEEE Intelligent Systems 15*(1), 8–17.

Heider, F. and M. Simmel (1944). An experimental study of apparent behavior. *Amer. J. Psych. 57*(2), 243–259.

Henriques, G. (2024, 6 March). Chatbots do not hallucinate, they confabulate. *Psych. Today*.

Hill, R. K. (2016). What an algorithm is. *Philosophy and Technology 29*, 35–59.

Hsu, T. and S. L. Myers (2023, 8 April). Can we no longer believe anything we see? *New York Times*.

Humphrey, N. (1976). The social function of intellect. In P. Bateson and R. Hinde (Eds.), *Growing Points in Ethology*, pp. 303–317. Cambridge Univ. Press.

Hutson, M. (2023, 16 May). Can we stop runaway AI? *The New Yorker*.

Jackson, F. (1986). What Mary didn't know. *J. Phil. 83*, 291–295.

Jefferson, G. (1949, 25 June). The mind of mechanical man. *Brit. Med. J. 1*(4616), 1105–1110.

Johnson-Laird, P. N. and M. Ragni (2023). What should replace the Turing test? *Intelligent Computing 2*. Article 0064.

Jordan, M. I. (2019). Artificial Intelligence—The revolution hasn't happened yet. *Harvard Data Science Review 1*(1).

Kabir, S., D. N. Udo-Imeh, et al. (2023). Who answers it better? An in-depth analysis of ChatGPT and Stack Overflow answers to software engineering questions. https://tinyurl.com/kabiretal2023.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Strauss and Giroux.

Kambhampati, S. (2023, 12 Sep.). Can LLMs really reason and plan? *Blog@CACM*. https://tinyurl.com/kambhampati2023.

Katz, D. M., M. J. Bommarito, et al. (2024). GPT-4 passes the bar exam. *Phil. Trans. Royal Soc. A 382*(2270).

Keller, H. (1905). *The Story of My Life*. Garden City, NY: Doubleday, 1954.

Kocijan, V., E. Davis, et al. (2023). The defeat of the Winograd Schema Challenge. *Artif. Intell. 235*, 103971.

Kolata, G. (2024, 17 Nov.). A.I. chatbots defeated doctors at diagnosing illness. *New York Times*.

Kolbert, E. (2023, 11 Sep.). Talk to me. *The New Yorker*, 44–53.

Kolbert, E. (2024, 3 October). Savvy in the grass. *New York Review of Books 71*(16), 6, 8.

Kriegel, U. (2024). Knowledge-by-acquaintance first. *Phil. & Phen. Res. 109*, 458–477.

Kung, T. H., M. Cheatham, et al. (2023). Performance of ChatGPT on USMLE. *PLOS Digital Health 2*(2), e0000198.

Labatut, B. (2023). *The MANIAC*. New York: Penguin.

Lakoff, G. and M. Johnson (1980). Conceptual metaphor in everyday language. *J. Phil. 77*(8), 453–486.

Lammens, J. (2024, 7 April). RAGs to ORGs: Or, how to make AI applications more transparent, interpretable, and accountable. https://tinyurl.com/2fj28e6o.

Landgrebe, J. and B. Smith (2021). Making AI meaningful again. *Synthese 198*, 2061–2081.

Landgrebe, J. and B. Smith (2023). *Why Machines Will Never Rule the World*. New York: Routledge.

Leibniz, G. W. (1714). *The Principles of Philosophy Known as Monadology*. J. Bennett (ed.) (2007), https://tinyurl.com/leibniz1714.

Lepore, J. (2024, 7 Oct.). The chit-chatbot. *The New Yorker*, 12–16.

Levesque, H. J. (2017). *Common Sense, the Turing Test, and the Quest for Real AI*. Cambridge, MA: MIT Press.

Levinstein, B. (2023a, 30 Jan.). A conceptual guide to transformers: Part I. *Substack*. https://tinyurl.com/levinstein2023PartI.

Levinstein, B. (2023b, 21 Feb.). What's going on under the hood of LLMs. *Substack*. https://tinyurl.com/levinstein2023PartV.

Levy, M. G. (2023, 1 Feb.). Machines learn better if we teach them the basics. *Quanta*.

Lewis, C. S. (1938). *Out of the Silent Planet*. New York: Macmillan, 1965.

Lewis, C. S. (1944). *Perelandra: A Novel*. New York: Macmillan, 1965.

Lewis, C. S. (1946). *That Hideous Strength*. New York: Macmillan, 1965.

Li, J., Y. Kementchedjhieva, et al. (2024). Do vision and language models share concepts? *Trans. Assoc. Comp. Ling. 12*, 1232–1249.

Lindstedt, J. (2006). Native Esperanto as a test case for natural language. In M. Suominen, A. Arppe, et al. (Eds.), *A Man of Measure*, pp. 47–55. Turku: Ling. Assoc. Finland.

Lohr, S. (2024, 23 July). When it comes to math, A.I. is dumb. *New York Times*.

Long, A. (2023, Jan.). On the ability to enjoy strawberries and cream with Alan Turing: An analysis of the imitation game. https://tinyurl.com/amandalong2023.

Lu, Y. (2024, 10 Apr.). Now hiring: Sophisticated (but part-time) chatbot tutors. *New York Times*.

Luccioni, S. and G. Marcus (2023, 17 Apr.). Stop treating AI models like people. *Marcus on AI*. https://tinyurl.com/luccioni-marcus20230417.

Macdonald, M. and M. Scriven (1954). The language of fiction. *Proc. Aristotelian Soc., Supp. Vol. 28*, 165–196.

MacFarquhar, L. (2007, 12 Feb.). Two heads: A marriage devoted to the mind-body problem. *The New Yorker*, 58–69.

Mann, T. (1924). *The Magic Mountain*. New York: Vintage Books, 1995. J.E. Woods (trans.).

Marcus, G. (2020, 19 Feb.). The next decade in AI: Four steps towards robust artificial intelligence. https://arxiv.org/abs/2002.06177.

Marcus, G. (2023a, 22 Sep.). Elegant and powerful new result that seriously undermines large language models. *Marcus on AI*. https://tinyurl.com/marcus20230922.

Marcus, G. (2023b, 19 Oct.). 'Math is hard'—if you are an LLM—and why that matters. *Marcus on AI*. https://tinyurl.com/marcus20231019.

Marcus, G. (2023c, 29 Oct.). What "game over" for the latest paradigm in AI might look like. *Marcus on AI*. https://tinyurl.com/marcus20231029.

Marcus, G. (2024a, 28 July). AlphaProof, AlphaGeometry, ChatGPT, and why the future of AI is neurosymbolic. *Marcus on AI*. https://tinyurl.com/marcus20240728.

Marcus, G. (2024b, 28 Nov.). ChatGPT, at age two. *Marcus on AI*. https://tinyurl.com/marcus20241128.

Marcus, G. (2024c, 9 Nov.). CONFIRMED: LLMs have indeed reached a point of diminishing returns. *Marcus on AI*. https://tinyurl.com/marcus20241109.

Marcus, G. (2024d, 5 Mar.). Covert racism in LLMs. *Marcus on AI*. https://tinyurl.com/marcus20240305.

Marcus, G. (2024e, 5 Feb.). Deconstructing Geoffrey Hinton's weakest argument. *Marcus on AI*. https://tinyurl.com/marcus20240205.

Marcus, G. (2024f, 13 Apr.). Evidence that LLMs are reaching a point of diminishing returns—and what that might mean. *Marcus on AI*. https://tinyurl.com/marcus20241413.

Marcus, G. (2024g, 21 Apr.). Humans versus machines: The hallucination edition. *Marcus on AI*. https://tinyurl.com/marcus20240421.

Marcus, G. (2024h, 30 June). Marcus goes gaga over Gates clip. *Marcus on AI*. https://tinyurl.com/marcus20240630.

Marcus, G. (2024i, 29 Feb.). Serious medical error from Perplexity's chatbot. *Marcus on AI*. https://tinyurl.com/marcus20240229.

Marcus, G. (2024j, 10 Feb.). Seven reasons why the world should say No to Sam Altman. *Marcus on AI*. https://tinyurl.com/marcus20240210.

Marcus, G. (2024k, 12 Apr.). Superhuman AGI is not nigh. *Marcus on AI*. https://tinyurl.com/marcus20240412.

Marcus, G. (2024l, 8 Feb.). There must be some misunderstanding. *Marcus on AI*. https://tinyurl.com/marcus20240208.

Marcus, G. and E. Davis (2019). *Rebooting AI.* New York: Vintage Books.

Marcus, G. and E. Davis (2023, 17 Feb.). How not to test GPT-3. *Marcus on AI*. https://tinyurl.com/marcus-davis20230217.

Martins, J. P. and S. C. Shapiro (1988). A model for belief revision. *Artif. Intell. 35*(1), 25–79.

Mays, W. (1952, April). Can machines think? *Philosophy 27*(101), 148–162.

McCarthy, J. and P. J. Hayes (1969). Some philosophical problems from the standpoint of Artificial Intelligence. In B. Meltzer and D. Michie (Eds.), *Machine Intelligence 4*. Edinburgh Univ. Press.

McGinn, C. (1999). Can we ever understand consciousness? *New York Review of Books 46*(10), 44–48.

Menabrea, L. F. and A. A. Lovelace (1843, Sep.). Sketch of the analytical engine invented by Charles Babbage, esq. *Scientific Memoirs 3*, 666–731. A.A. Lovelace (trans.), https://tinyurl.com/menabrea1843. Lovelace's notes at https://tinyurl.com/lovelace1843.

Metz, C. (2023a, 25 Sep.). The secret ingredient of ChatGPT is human advice. *New York Times*.

Metz, C. (2023b, 29 Mar.). What makes A.I. chatbots go wrong? *New York Times*.

Metz, C. (2024a, 23 Sep.). Is math the path to chatbots that don't make stuff up? *New York Times*.

Metz, C. (2024b, 13 Feb.). OpenAI gives ChatGPT a better 'memory'. *New York Times*.

Michael, J. (2020, 23 July). To dissect an octopus. https://tinyurl.com/jmichael20200723.

Michie, D. (1971). Formation and execution of plans by machine. In N. Findler and B. Meltzer (Eds.), *Artificial Intelligence and Heuristic Programming*, pp. 101–124. New York: American Elsevier.

Mickle, T. (2024, 9 Sep.). Can Apple's iPhone pass the A.I. test? *New York Times*.

Mirzadeh, I., K. Alizadeh, et al. (2024, Oct.). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. https://tinyurl.com/mirzadeh-etal2024.

Mitchell, M. (2023). How do we know how smart AI systems are? *Science 381*(6654).

Mitchell, M. et al. (2023, 11 Dec.). Comparing humans, GPT-4, and GPT-4V on abstraction and reasoning tasks. https://arxiv.org/pdf/2311.09247.

Mols, B. (2023, 27 Apr.). Artificial intelligence still can't form concepts. *Commun. ACM*. https://tinyurl.com/mols20230427.

Moravec, H. (1998). When will computer hardware match the human brain? *J. Evol. Tech. 1*(1).

Morris, C. (1938). *Foundations of the Theory of Signs*. Chicago: University of Chicago Press.

Nazaryan, A. (2024, 3 Jan.). Could monkeys really type all of shakespeare? *New York Times*. https://tinyurl.com/nazaryan2024.

Neely, A. (2024). Why AI is doomed to fail the musical Turing Test. YouTube video. Viewed 8 May 2024 at https://tinyurl.com/neely2024.

Neilson, W. A., T. A. Knott, and P. W. Carhart (Eds.) (1958). *Webster's New International Dictionary of the English Language, 2nd Edition, Unabridged*. Springfield, MA: G.& C. Merriam.

Newell, A., J. Shaw, and H. A. Simon (1958). Elements of a theory of human problem solving. *Psychological Review 65*(3), 151–166.

Newport, C. (2023, 13 Apr.). What kind of mind does ChatGPT have? *The New Yorker*. https://tinyurl.com/newport20230413.

Newport, C. (2024, 15 Mar.). Can an A.I. make plans? *The New Yorker*.

O'Connor, M. T. (2024). In the craftsman's garden: AI, Alan Turing, and Stanley Cavell. *Minds and Machines 34*(22).

OpenAI (2023, 23 Mar.). GPT-4 system card. https://tinyurl.com/openai20230323.

Pardo, M. A., K. Fristrup, et al. (2024, 10 June). African elephants address one another with individually specific name-like calls. *Nat. Ecol. Evol.*.

Parnas, D. L. (2017). The real risks of artificial intelligence. *Communications of the ACM 60*(10), 27–31.

Pennycook, G., J. A. Cheyne, et al. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition 123*, 335–346.

Perconti, P. and A. Plebe (2023). Do machines really understand meaning? (again). *J. Artif. Intell. & Consciousness 10*(1), 181–206.

Perry, J. (1979). The problem of the essential indexical. *Noûs 132*(1), 3–21.

Perry, Jr., W. G. (1970). *Forms of Intellectual and Ethical Development in the College Years: A Scheme*. New York: Holt, Rinehart and Winston.

Perry, Jr., W. G. (1981). Cognitive and ethical growth. In A. C. et al. (Ed.), *The Modern American College*, pp. 76–116. San Francisco: Jossey-Bass. https://tinyurl.com/perry81.

Pettit, P. (1991, October). Realism and response-dependence. *Mind 100*(4), 587–626.

Piccinini, G. (2016). The computational theory of cognition. In V. Müller (Ed.), *Fundamental Issues of Artificial Intelligence*, pp. 203–221. Berlin: Springer.

Posner, R. (1992). Origins and development of contemporary syntactics. *Languages of Design 1*, 37–50.

Proudfoot, D. (2005). A new interpretation of the Turing test. *The Rutherford Journal 1*.

Proudfoot, D. (2013). Rethinking Turing's test. *J. Phil. 110*(7), 391–411.

Proudfoot, D. (2017). The Turing test—from every angle. In B. J. Copeland et al. (Eds.), *The Turing Guide*, pp. 287–300. Oxford Univ. Press.

Proudfoot, D. (2024). Wittgenstein and Turing on AI. In A. C. Helliwell et al. (Eds.), *Wittgenstein and Artificial Intelligence, Vol. 1*, pp. 17–37. London: Anthem Press.

qFiasco, F. (2018). Book review [of Kasparov & Greengard, *Deep Thinking*]. *Artif. Intell. 260*, 36–41.

Quilty-Dunn, J. et al. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behav. Brain Sci. 46:e261*.

Quine, W. V. O. (1987). Universal library. In *Quiddities*, pp. 223–235. Harvard Univ. Press.

Raiola, R. (2023). ChatGPT, can you tell me a story? *Commun. ACM 66*(5), 102–104.

Rapaport, W. J. (1984). Can philosophy solve its own problems? *The [SUNY] News 13*, F2–F3. https://tinyurl.com/rapaport1984sunynews.

Rapaport, W. J. (1985). To be and not to be. *Noûs 19*(2), 255–271.

Rapaport, W. J. (1986a, Summer). Philosophy, Artificial Intelligence, and the Chinese-room argument. *Abacus: The Magazine for the Computer Professional 3*, 6–17. http://www.cse.buffalo.edu/~rapaport/Papers/abacus.pdf.

Rapaport, W. J. (1986b). Searle's experiments with thought. *Phil. Sci. 53*, 271–279.

Rapaport, W. J. (1988). Syntactic semantics. In J. H. Fetzer (Ed.), *Aspects of Artificial Intelligence*, pp. 81–131. Dordrecht: Kluwer.

Rapaport, W. J. (1995). Understanding understanding: Syntactic semantics and computational cognition. In J. E. Tomberlin (Ed.), *AI, Connectionism, and Philosophical Psychology (Philosophical Perspectives, Vol. 9)*, pp. 49–88. Atascadero, CA: Ridgeview.

Rapaport, W. J. (1998). How minds can be computational systems. *J. Exp. Theor. Artif. Intell. 10*, 403–419.

Rapaport, W. J. (2000). How to pass a Turing test. *J. Logic, Lang. Info. 9*(4), 467–490.

Rapaport, W. J. (2002). Holism, conceptual-role semantics, and syntactic semantics. *Minds and Machines 12*(1), 3–59.

Rapaport, W. J. (2003a). What did you mean by that? Misunderstanding, negotiation, and syntactic semantics. *Minds and Machines 13*(3), 397–427.

Rapaport, W. J. (2003b). What is the 'context' for contextual vocabulary acquisition? In P. P. Slezak (Ed.), *Proc. 4th Int'l. Conf. Cog. Sci./7th Australasian Soc. Cog. Sci. Conf.*, pp. Vol. 2, pp. 547–552. Univ. New South Wales. https://tinyurl.com/rapaport2003context.

Rapaport, W. J. (2005). In defense of contextual vocabulary acquisition. In A. Dey, B. Kokinov, et al. (Eds.), *Modeling and Using Context*, pp. 396–409. Springer-Verlag Lecture Notes Artif. Intel. 3554.

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

Rapaport, W. J. (2011). Yes, she was! Reply to Ford's "Helen Keller was never in a Chinese room". *Minds and Machines 21*(1), 3–17.

Rapaport, W. J. (2015, 7 Jan.). A history of the sentence 'Buffalo buffalo buffalo buffalo buffalo.'. https://tinyurl.com/rapaportbuffalobuffalo.

Rapaport, W. J. (2017). Semantics as syntax. *Amer. Phil. Assoc. Newsletter Phil. Computers 17*(1), 2–11.

Rapaport, W. J. (2018, 22 Dec.). William Perry's scheme of intellectual and ethical development. https://cse.buffalo.edu/~rapaport/perry-positions.html.

Rapaport, W. J. (2020). What is Artificial Intelligence? *J. Artif. General Intell. 11*(2), 52–56.

Rapaport, W. J. (2023). *Philosophy of Computer Science*. Wiley-Blackwell.

Rapaport, W. J. (2024a). Is artificial general intelligence impossible? *Cosmos + Taxis 12*(5+6), 5–22.

Rapaport, W. J. (2024b). Yes, AI *can* match human intelligence. https://cse.buffalo.edu/~rapaport/Papers/aidebate.pdf. Draft of book chapter in progress.

Rapaport, W. J. and M. W. Kibby (2007). Contextual vocabulary acquisition as computational philosophy and as philosophical computation. *J. Exp. Theor. Artif. Intell. 19*(1), 1–17.

Rapaport, W. J. and M. W. Kibby (2014). Contextual vocabulary acquisition: From algorithm to curriculum. In A. Palma (Ed.), *Castañeda and His Guises*, pp. 107–150. Berlin: de Gruyter.

Rapaport, W. J., E. M. Segal, et al. (1989). Cognitive and computer systems for understanding narrative text. *SUNY Buffalo Center for Cog. Sci. Tech. Rept. 89-07*. https://tinyurl.com/rapaport-etal1989.

Rapaport, W. J., S. C. Shapiro, and J. M. Wiebe (1997). Quasi-indexicals and knowledge reports. *Cog. Sci. 21*, 63–107.

Rescher, N. (1978). Philosophical disagreement. *Rev. Metaphysics 32*(2), 217–251.

Rogers, Jr., H. (1959). The present theory of Turing machine computability. *J. Soc. Industrial Applied Math. 7*(1), 114–130.

Roose, K. (2024a, 30 Aug.). How do you change a chatbot's mind? *New York Times*.

Roose, K. (2024b, 9 May). Meet my A.I. friends. *New York Times*.

Roschelle, J. (2023, 30 July). Stop judging AI using human exams. *Blog@CACM*. https://tinyurl.com/roschelle20230730.

Rothman, J. (2023, 20 Nov.). Metamorphosis (why the godfather of A.I. fears what he's built). *The New Yorker*, 28–39.

Rothman, J. (2024, 27 Aug.). What does it really mean to learn? *The New Yorker*.

Rowling, J. (1999). *Harry Potter and the Chamber of Secrets*. New York: Scholastic.

Saba, W. S. (2023, 23 May). How did scientists succumb to Aunt Edna? The dangers of a superintelligent AI is fiction. *Blog@CACM*. https://tinyurl.com/wssaba2023.

Sablé-Meyer, M., J. Fagot, et al. (2021). Sensitivity to geometric shape regularity in humans and baboons. *PNAS 118*(16).

Sahlgren, M. and F. Carlsson (2021, 7 Sep.). The singleton fallacy: Why current critiques of language models miss the point. *Frontiers Artif. Intell. 4*(682578).

Savage, N. (2024a). Beyond Turing: Testing LLMs for intelligence. *Commun. ACM 67*(9), 10–12.

Savage, N. (2024b). Can ChatGPT learn Chinese or Swahili? *Commun. ACM 67*(5), 26–31.

Saygin, A. P., I. Cicekli, and V. Akman (2000). Turing test: 50 years later. *Minds and Machines 10*, 463–518.

Schaeffer, R. et al. (2023). Are emergent abilities of large language models a mirage? In *ICML 2023 Workshop: Challenges of Deploying Generative AI*. https://tinyurl.com/schaeffer-etal2023.

Schank, R. C. (n.d.). Speculations on understanding, mechanical and human. Unpublished talk at SUNY Buffalo(?). Copy available upon request from rapaport@buffalo.edu.

Scholes, R. (1968). *Elements of Fiction*. New York: Oxford University Press.

Schulte, P. (2023). *Mental Content*. Cambridge, UK: Cambridge Univ. Press.

Schwitzgebel, E. et al. (2023). Creating a large language model of a philosopher. *Mind & Language 39*(2), 237–259.

Seabrook, J. (2019, 14 Oct.). The next word. *The New Yorker*, 52–63.

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci. 3*, 417–457.

Searle, J. R. (1982, 29 Apr.). The myth of the computer. *New York Review of Books*, 3–6.

Searle, J. R. (1986). Turing the Chinese room. *Proc. Heraclitean Soc. 11*.

Segal, E. M. (1995). A cognitive-phenomenological theory of fictional narrative. In J. F. Duchan et al. (Eds.), *Deixis in Narrative*, pp. 61–78. Erlbaum.

Sejnowski, T. J. (2023). Large language models and the reverse Turing test. *Neural Computation 35*(3), 309–342.

Sells, P. (1987). Aspects of logophoricity. *Linguistic Inquiry 18*, 445–479.

Shanahan, M. (2024). Talking about large language models. *Commun. ACM 67*(2), 68–79.

Shapiro, S. C. (1992). Artificial Intelligence. In S. C. Shapiro (Ed.), *Encyc. Artif. Intell., 2nd Ed.*, pp. 54–57. New York: Wiley.

Shapiro, S. C. and W. J. Rapaport (1991). Models and minds: Knowledge representation for natural-language competence. In R. Cummins and J. Pollock (Eds.), *Philosophy and AI*, pp. 215–259. Cambridge, MA: MIT Press.

Sieg, W. (2008). On computability. In A. Irvine (Ed.), *Philosophy of Mathematics*, pp. 525–621. Oxford: Elsevier.

Simon, H. A. (1996). Computational theories of cognition. In W. O'Donohue and R. F. Kitchener (Eds.), *The Philosophy of Psychology*, pp. 160–172. London: SAGE Publications. Page ref. to https://tinyurl.com/simon1996.

Simon, H. A. and A. Newell (1958, Jan.-Feb.). Heuristic problem solving. *Operations Research 6*(1), 1–10.

Singer, N. (2023, 6 Feb.). At this school, computer science class now includes critiquing chatbots. *New York Times*.

Singhal, K., S. Azizi, et al. (2023). Large language models encode clinical knowledge. *Nature 620*, 172–180.

Sloman, A. (2010, 7 January). Why symbol-grounding is both impossible and unnecessary, and why symbol-tethering based on theory-tethering is more powerful anyway. https://www.cs.bham.ac.uk/research/projects/cogaff/misc/talks/models.pdf.

Smith, A. (1776). *Wealth of Nations*. https://tinyurl.com/yc95pvk4. For "On the Division of Labor" (Book I, Ch. I), link to https://tinyurl.com/smith1776bk1ch1. For Book II, Ch. I, link to https://tinyurl.com/smith1776ch01c-2.

Smith, B. C. (2019). *The Promise of Artificial Intelligence*. Cambridge, MA: MIT Press.

Søgaard, A. (2023). Grounding the vector space of an octopus. *Minds and Machines 33*, 33–54.

Srihari, R. K. (1991a). Extracting visual information from text. PhD Diss. Dept. Comp. Sci., SUNY Buffalo.

Srihari, R. K. (1991b). Piction: A system that uses captions to label human faces in newspaper photographs. In *Proc. 9th Nat'l. Conf. Artif. Intell.*, pp. 80–85.

Srihari, R. K. and W. J. Rapaport (1989). Extracting visual information from text. In *Proc. 11th Annual Conf. Cog. Sci. Soc.*, pp. 364–371. Erlbaum.

Srihari, R. K. and W. J. Rapaport (1990). Combining linguistic and pictorial information. In D. Kumar (Ed.), *Current Trends in SNePS—Semantic Network Processing System*, pp. 85–96. Berlin: Springer-Verlag Lecture Notes Artif. Intell. 437.

Stein, D. K. (1984). Lady Lovelace's notes. *Victorian Studies 28*(1), 33–67.

Stich, S. and I. Ravenscroft (1994). What is folk psychology? *Cognition 50*(1–3), 447–468. Page ref. to https://tinyurl.com/stich-ravenscroft1994.

Stria, I. (2015). Esperanto as a natural language. *Language. Communication. Information 10*, 32–42.

Tamayo-Sarver, J. (2023, 5 Apr.). I'm an ER doctor: Here's what I found when I asked ChatGPT to diagnose my patients. *Inflect Health*.

Thorpe, C. (1989, September). *Computing Reviews*. Cover quotation, dated 22 August 1989.

Titus, L. M. (2023). Does ChatGPT have semantic understanding? *Cog. Sys. Res. 83*(101174).

Trott, S., C. Jones, et al. (2023). Do large language models know what humans know? *Cog. Sci. 47*(e13309).

Turing, A. M. (1936). On computable numbers. *Proc. London Math. Soc., Ser. 2, 42*, 230–265.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind 59*(236), 433–460.

Turing, A. M. et al. (1952). Can automatic calculating machines be said to think? In B. J. Copeland (Ed.), *The Essential Turing*, pp. 487–506. Oxford Univ. Press, 2004.

Turney, P. D. and P. Pantel (2010). From frequency to meaning. *J. Artif. Intell. Res. 37*, 141–188.

van Dis, E. A. et al. (2023, 9 Feb.). ChatGPT: Five priorities for research. *Nature 614*, 224–226.

van Inwagen, P. (1984, Fall). On computer simulation of the presence of a thinking being. *Syracuse Scholar*, 15–21.

Vaswani, A., N. Shazeer, et al. (2017). Attention is all you need. In *31st Conf. Neural Info. Processing Systems (NIPS 2017)*.

von Hippel, P. T. (2023, 4 Jan.). ChatGPT is not ready to teach geometry (yet). *Education Next*.

Vong, W. K., W. Wang, et al. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science 383*(6682), 504–511.

Waldo, J. and S. Boussard (2024, 6 Dec.). GPTs and hallucination. *Commun. ACM*.

Waxman, O. B. (2024). AI in the classroom. In A. Vered and P. Weintraub (Eds.), *Artificial Intelligence: A New Age of Possibilities*, pp. 44–47. Time [Magazine].

Wei, J. et al. (2022). Emergent abilities of large language models. *Trans. Machine Learning Res.*. https://arxiv.org/pdf/2206.07682.

Weil, E. (2023, 1 Mar.). You are not a parrot. *New York Magazine*.

Weiser, B. (2023a, 22 June). ChatGPT lawyers are ordered to consider seeking forgiveness. *New York Times*.

Weiser, B. (2023b, 27 May). Here's what happens when your lawyer uses ChatGPT. *New York Times*.

Weiser, B. and N. Schweber (2023, 8 June). The ChatGPT lawyer explains himself. *New York Times*.

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM 9*(1), 36–45. Reprinted in *CACM* 26(1): 23–28; be sure to read the correspondence between Weizenbaum and Arbib on p. 28 of the reprint.

Weizenbaum, J. (1967). Contextual understanding by computers. *Commun. ACM 10*(8), 474–480.

Weizenbaum, J. (1976). *Computer Power and Human Reason*. New York: W.H. Freeman.

Whang, O. (2024, 30 Apr.). From baby talk to baby A.I. research. *New York Times*.

Wittgenstein, L. (1933–1934). *The Blue and Brown Books*. Oxford: Basil Blackwell, 1964.

Wittgenstein, L. (1953). *Philosophical Investigations, 4th ed.* Wiley-Blackwell, 2009. Trans. G.E.M. Anscombe et al.

Włodarczyk, M. (2023, 13 March). #ChatGPT_dialogues. https://tinyurl.com/wlodarczyk2023.

Wolfram, S. (2023). *What Is ChatGPT Doing... and Why Does It Work?* Wolfram Media. wolfr.am/SW-ChatGPT and wolfr.am/ChatGPT-WA.

Wolman, D. (2012, 15 March). A tale of two halves. *Nature 483*, 260–263.

Zhang, Y. and K. Davidson (2024). Interpreting referential noun phrases in belief reports. *Glossa Psycholinguistics 3(1)*(13).

Zhou, J. et al. (2023). Synthetic lies. In *Proc. 2023 CHI Conf. on Human Factors in Computing Systems*, pp. 1–20. ACM. Article 436.