Looking the Part: The AI Zombie Problem and the Anti-Turing Test

Julien da Silva †*
† McGill University

Abstract

This paper addresses a philosophical and practical issue with the human tendency to anthropomorphise artificial intelligence. Specifically, I argue that there is a need to regulate the widespread adoption and societal integration of human-like AI, given that there will be substantive, irresolvable disagreement over their moral status. My aim in this paper is to propose a viable regulatory intervention that will avoid the social and political confusion that would be generated by human-like AI. I favour a novel solution called the Anti-Turing Test: a modified Turing Test that reliably distinguishes problematic human-like AI from others suitable for social integration.

Keywords: Anthropomorphism, Turing Test, Social Robot, Philosophical Zombie

1. Introduction

A human-like artificial intelligence (AI) approximates the physical and intellectual capacities of a human being. This encompasses a broad range of systems, but in this paper, I’m particularly interested in human-like social robots and, to a lesser extent, conversational agents (i.e., chatbots). Social robotics is a developing field at the intersection of artificial intelligence and robotics research. A social robot possesses characteristics such as communication through social cues, displaying emotion, and adaptive behaviour, to facilitate social interaction with humans [1]. Presently, many of these robots are confined to medical applications, recreation, or laboratory settings. But progress in this domain is not expected to slow, and social robots are predicted to integrate more deeply into our daily lives, for example, as service workers. Conversational agents, by contrast, are more often limited to existence as software on a computer or server. This kind of AI is specifically designed for natural conversation through a user interface. As of recently, ChatGPT has been in the spotlight for being one of the most advanced implementations of this technology. [2]

This paper addresses a philosophical and practical issue with the human tendency to anthropomorphise artificial intelligence. Specifically, I argue that there is a need to regulate the widespread adoption and integration of human-like AI, given that there will be substantive, irresolvable disagreement over their moral status. For this reason, my aim in this paper is also to propose regulatory interventions that will avoid the social and political confusion that would be generated by human-like AI. Here, I canvas several viable solutions, but I favour a novel proposal called the Anti-Turing Test: a modified Turing Test that reliably distinguishes problematic human-like AI from others suitable for social integration.

I divide this paper into four sections. First, I introduce a recent case study in which a conversational AI is perceived to have human-like sentience. In Section Two, I introduce Susan Schneider’s concept of an AI zombie, a human-like AI with all the outward properties of consciousness without being conscious. In Section Three, I explain the problems with mistaking AI zombies for conscious beings. Specifically, I expound on the issue with treating AI zombies with human-like moral consideration and the social and political confusion that would cause. Finally, in Section Four, I survey potential solutions to these problems. Here, I examine whether policy informed by the Turing Test and its derivatives could be a viable avenue for AI zombie regulation. Here, I propose an original solution called the Anti-Turing
Test, an empirical test meant to accurately delineate between AI that would cause confusion and those that would not.

2. LaMDA: First Contact

It is June of 2022. Blake Lemoine, a member of Google’s Responsible AI organization, opened his laptop to converse with an experimental piece of discursive software, the Language Model for Dialogue Applications, also known as LaMDA. The system was effectively an advanced conversational interface specifically designed for natural dialogue. Whereas Lemoine’s task was to determine whether it would express discriminatory attitudes, his conversations with the program took a turn when it seemed to express philosophical beliefs about the nature of its own existence, the soul, personhood, justice, and human rights. Through his conversations, he became convinced LaMDA possessed a consciousness of its own and wanted the world to know. In the aftermath of the media coverage, the consensus among philosophers and technologists alike was against Lemoine. LaMDA was not conscious; it was simply an impressive chatbot.

Lemoine’s reaction is an example of the human propensity to project our emotions and other mental states onto animals and inanimate objects. Psychological literature abounds about anthropomorphism; we speak about things as if they have intentions (“this car is trying to kill me!”), we ascribe human feelings to our pets, and we are even able to conjure complex stories from the movement of shapes on a screen [3], [4], [5].

There is a similarly rich investigation in the ethics of technology examining the effects of our anthropomorphic tendency on our interactions with robots and AI. Forward-looking thinkers like Darling [1] and Turkle [6] predict that social robots will enter our homes and social spheres. Darling, in particular, recounts the already revealing ways we project human-like characteristics onto machines, from the mundanity of naming our Roombas, to more sombre examples, like soldiers mourning the “death” of a land mine-detecting robot.

While these examples seem innocent or otherwise non-threatening, Lemoine’s case is different. His conversation with LaMDA convinced him of its sentence. In his own words, “I know a person when I talk to it” [7]. He claimed not just that it understood moral concepts, but also that it deserved to be treated with the dignity and respect of a human being. In this way, Lemoine is making a much less innocuous and much more radical claim about the moral status of the machine. At least, according to Lemoine, he was dealing with another, full-fledged person.

The LaMDA case foreshadows an immediately relevant problem facing AI ethicists, technologists, and policymakers alike. Chat software and social robots alike will be convincingly human in their speech patterns and behaviour. In a widespread adoption scenario, ordinary people will, naturally and reasonably, take the social cues, expressions of interiority, and fluent speech of these AI systems as indicators of consciousness.

3. Anthropomorphism and Moral Status

The race toward human-like AI is garnering philosophical attention. These systems are forcing us to re-examine whether our cognitive faculties (creativity, critical thinking, language use, etc.) are uniquely human. The end goal, I suspect, is to unify each of these capabilities into one system. The thought goes that it would then have a general intelligence to rival our own. Intelligence aside, however, it is not obvious that instantiating one or even multiple of these capabilities in an AI system will produce another of our human faculties: consciousness. This is because we neither have a philosophical nor neuroscientific theory of consciousness complete enough to identify what thing(s) consciousness supervenes on.

This is a bigger problem than it first appears. There are reasons to suspect that it will be possible to create an artificial agent capable of moving and talking like a human, thereby exhibiting all the qualities of consciousness to an ordinary observer but lacking it entirely [6, Appendix A]. Additionally, until we can isolate the minimal neural correlates for consciousness, we should expect to create AI zombies: an AI zombie is a human-like AI that
convincingly possesses the outward signs of consciousness without actually being conscious.

To recap:

**Claim 1:** It is possible to mistakenly perceive human-like consciousness in a profoundly human-like machine (or: AI zombies are possible).

**Claim 2:** Until consciousness is well understood, it is more likely that we will make AI zombies than conscious AI.

Sophisticated conversational AI are already potential candidates for AI zombies. Specifically, these systems exploit a quirk in our cognition. We are accustomed to the idea that any instances of fluent language use are caused by an intelligent, thinking being. This is a narrow application of Dennett's (1971) claim that when we observe some phenomenon, we can respond to it in one of two modes [9]. On the one hand, we can take a physical stance towards the phenomenon and rely on our knowledge of natural laws to respond and make predictions. Say, if I dropped a stone. But on the other hand, say in cases where we interact with others, we adopt an intentional stance, in which we treat the thing as an agent with intentions, goals, beliefs, and feelings; responding to language is facilitated by taking an intentional stance. But what about the language of a conversational AI? In a near future where social robots walk like humans and talk like humans, is it not natural for people to believe that they should respond like they are human, too?

Taking an intentional stance is a kind of anthropomorphism. In cases like these, people imagine an intelligent mind in the machine animating the words that they read and hear. However, to the extent that we treat AI beings like agents with fluent intentions and beliefs, we also project consciousness onto them and all the normative commitments that entails.

We are now able to spell the problem out directly. Generally, we take it that having human-like consciousness is a requirement for deserving human-like moral consideration [8, Appendix B]. So, conferring consciousness onto an AI zombie would be a grave error, as they would be unworthy of this kind of moral consideration. However, this error would also be likely given the high likelihood of AI zombies. So, given Claim 1 and Claim 2 together:

**Claim 3:** If human-like AI becomes widely adopted, then there would likely be widespread error about its human-like moral status.

By widely adopted, I mean available for ordinary people not only to purchase for personal use but also commonplace in social spaces. By human-like moral status, I mean the status conferred upon by our ability to think, feel, introspect, grasp meanings of moral concepts, to feel pain and pleasure, all of which stem from our human consciousness.

4. **Why Does This Matter?**

What I am describing might seem a bit far-fetched or like science fiction. For this reason, two objections to this project come to mind. First, perhaps I've been too hasty in the setup of this problem. Is AI consciousness even possible?

For one, the likelihood of AI consciousness is still an open debate; some philosophers like Wallach and Allen argue that “fully conscious artificial systems with complete human moral capability may remain forever in the realm of science fiction” [9, p. 8]. Others, like Bostrom [12] and Chalmers [13], believe otherwise.

Moreover, depending on one’s theoretical commitments in the philosophy of mind, the entire framing may be flawed from the start. For example, taking up a biological naturalist view in the style of John Searle and rejecting that software can instantiate anything like understanding, thoughts, consciousness, or otherwise, rules out the possibility of human-like moral consideration for AI ex hypothesi [14]. Nevertheless, biological naturalism is far from an uncontroversial view, and there are plausible competing views that do not rule out machine consciousness [15], [16], [Appendix C]. For these reasons, it will be prudent to move
forward in a theory-neutral manner, and that means leaving open at least the possibility of conscious AI.

Second, we could agree that treating an AI zombie with the same moral consideration as a human being would be a mistake. So what? Does this mistake have consequences, or if it does, then why should we care now, especially when human-like AI seem so distant?

There are reasons to start thinking about regulating future technology now. For instance, if there is one lesson technology ethicists and policymakers have learned with the rise of social media, algorithmic curation of online content, and the way these interact to create echo chambers and polarize societies, it is that new technology can be volatile and harmful when left to grow entirely unregulated, only for policy to retroactively pick up the pieces after the damage has already been done. If human-like AI have the potential for human-like moral consideration, then we should expect this technology to be equally disruptive as others, if not more. So, it would be prudent to act pre-emptively to ensure a smooth integration of human-like AI into the public sphere.

Another reason we should care about conscious AI is that things with human-like moral considerations also deserve human-like legal considerations. But if we forestall finding a way to identify conscious machines, there is a chance that by the time conscious AI arrive, we will not be prepared; they will have different rights and freedoms than humans. The opposite scenario seems just as wrong: this is because, on the other hand, it would be a waste of precious time and resources to draft policies and protections for human-like AI when they are not conscious in the first place, especially when there are flesh-and-blood human beings in less fortunate situations who should be the focus of our collective political resources. So, there is pressure to get this right, and one way to do that is to start to think about this problem soon. Additionally, Schneider underscores that robots are already being developed to perform maintenance on nuclear reactors, fight wars, and other dangerous tasks. If we ask a robot to perform these tasks on our behalf, we should be sure of its status, even if to avoid needlessly putting human-like beings in harm’s way [15, p. 440].

There are several other reasons we might care about conscious AI now, but here is just one more. Suppose AI zombies that behave and speak like humans emerge into the public sphere unregulated, and ordinary people reasonably conclude from appearances that they deserve human-like moral considerations. In that case, it follows that people will consider these robots blameworthy.

In the literature on moral responsibility, blameworthiness is typically distinguished from just being to blame. In our ordinary thought and talk, blaming something can mean pointing out the cause of some effect. For example, blaming your dog for digging holes in your backyard, or blaming the bad weather for your wet hair, means little else than describing the causal connections between one thing and the other. In these cases, holding your dog morally accountable for ruining your backyard or the rain for bad hair makes no sense.

In contrast, if an agent is blameworthy, then aside from being causally responsible for some action, the agent must also have a kind of “normative competence,” according to Gary Watson:

> A person’s status as a responsible agent requires not only the capacity to conform her desires and conduct to her deepest values ... but also the capacity to acquire the right values – that is, those we hold her responsible for having [16, p. 228].

If the picture I have described of AI zombies is right, then they would appear to possess all the necessary indicators of normative competence. But, to the extent that AI zombies lack consciousness and by extension, lack the inner world necessary for having desires, values, and understanding the meanings of moral terms, it is not obvious that they could possess the normative competence required for blameworthiness [17, Appendix D].

To bring this into focus, imagine if a human-like AI in the form of an autonomous soldier killed an innocent civilian. Who or what is considered morally blameworthy in this case? The
answer depends partly on whether the autonomous soldier is conscious. This will, in turn, inform how we treat the soldier: as an inanimate weapon of war used to commit a war crime or as a blameworthy agent in itself who committed a war crime. If the soldier were an AI zombie, then it would be appropriate to assign blame up the chain of command, identifying those people who authorized the use of a defective weapon and enacting justice that way. But if the AI zombie soldier were misidentified as conscious, we would dole out punishments to a being that cannot feel or think, which would be a failure to enact proper justice.

Worse still, all these scenarios will play out the world over in a scenario where human-like AI zombies become fixtures of public life. We should expect ordinary people with varying degrees of familiarity with AI to have profound and irresolvable disagreements about the moral status of the very machines they are engaging with daily. This confusion would cause damage to the social fabric and should be avoided. But how?

5. Solutions

Recall that our goal is to ensure that ordinary people do not misattribute consciousness to AI zombies. Regulatory oversight will likely be necessary to motivate a favourable outcome. Specifically, we should want policy that helps minimize opportunities for perceiving consciousness where there is not any. One place to start might be the Turing Test.

5.1. The Turing Test (& Derivatives) Informing Policy?

The Imitation Game, more commonly known as the Turing Test, was devised by Alan Turing in 1950 to test if a machine could exhibit intelligent behaviour indistinguishable from a human [20]. Originally conceived, the test comprised a human judge, a human contestant, and a machine contestant. The judge then engages in a natural language conversation with the human and the machine without knowing which is which. A machine is said to have passed the Turing Test just in case the judge could not reliably determine which of the contestants is the machine and which is the human. According to Turing, if a machine has passed the Turing Test, it would qualify as thinking. So it goes, perhaps it would make sense to introduce regulation along the lines of the outcomes of a Turing Test. That is, we could avoid mistaking AI zombies for bona fide conscious AI by subjecting all new human-like AI to a Turing Test-style scenario. If it passes, it would be suitable to confer on it a human-like moral status.

Not so fast, though. The Turing Test has sparked several discussions and criticisms in the literature, many of which Turing did not address. In the wake of these contributions, it has become clear that the Turing Test is not a suitable way to detect conscious machines. One way to see this is by asking the following question. Is passing the Turing Test meant to establish a necessary or a sufficient condition for thinking beings? On the one hand, it seems at least conceivable that highly intelligent, conscious beings can fail the Turing Test if they do not use language the same way we do. Given the test’s reliance on language to communicate intelligence, it seems like it would leave out cases like these, so passing the test cannot constitute a necessary condition. On the other hand, it could not establish a sufficient condition either. This was famously demonstrated by Ned Block’s thought experiment in which he imagined a creature that looks identical to a human being but responds to language by searching through a lookup table of coherent replies and selecting one [21]. Block claimed that this could pass the Turing Test but failed to instantiate intelligence (in our case, thoughts, understanding, meaning, etc.). For this reason, passing the Turing Test is not a sufficient condition for having thoughts either.

Others have tried to salvage the project Turing began by devising new, more robust trials for human-like AI to clear. One of the first original proposals by Bringsjord, Bello, and Ferrucci is called the Lovelace Test, inspired by Ada Lovelace’s claim that machines cannot create original things or have original thoughts [22]. An AI agent can pass the Lovelace Test if the designer of the agent cannot explain how it produced a given output and the output was not a fluke.
Mark Riedl thinks no agent could pass this test [23]. Specifically, he thinks that any designer capable of producing an AI agent can, with sufficient time, also explain how the agent produced the output. If it were unbeatable, then clearly it would serve as a poor foundation for policy regulating human-like AI. But I think if discussions around the AI “Black Box Problem” have taught us anything, it is that even the very designers of the AI tools do not entirely understand how the system arrives at all its outputs [24]. So, on the contrary, I think very many AI systems could pass this test. I just do not think that an AI’s status as creative, intelligent, conscious, or otherwise should depend on whether some software engineers can explain how their system works, nor should our policy depend on this either. Riedl iterates on the test with his Lovelace 2.0 test for detecting creativity in artificial agents but concedes that it’s only capable of comparing agents relative to other agents; “passing” this test is neither necessary nor sufficient for having intelligence or thought.

Moving away from Turing Test-style proposals, Schneider offers both the AI Consciousness Test (ACT) and the Chip Test, which focus on ways to specifically distinguish between two identical human-like AI wherein one possesses phenomenal consciousness, and the other does not. In this paper, I focus on the ACT because it is the most actionable in terms of policy. The test is set up so that the robot must be able to describe, quickly and readily, certain situations that seem to depend on having a conception of phenomenal consciousness, such as reincarnation, out-of-body experiences, body switching, and more. For Schneider:

*These scenarios would be exceedingly difficult to comprehend for an entity that had no conscious experience whatsoever. It would be like expecting someone who is completely deaf from birth to appreciate a Beethoven symphony* [15, p. 443].

This test targets much more directly thing at issue, namely distinguishing AI zombies and conscious AI. Moreover, passing the ACT is a sufficient condition for consciousness.

However, Udell and Schwitzgebel remain unconvinced of the claim that the ACT establishes a sufficient condition for consciousness in AI [25]. Specifically, they think that it should always be possible to give at least one explanation for the system’s replies entirely in terms of its “architectural makeup” – the way it works mechanically, its software, and so on. If its response can be wholly explained by the system’s architecture, then it competes with the explanation that its response can be explained by its consciousness. Nevertheless, they see value in her test as a useful heuristic method for detecting phenomenally conscious AI paired with an “enhanced Turing Test,” which they believe would be an especially difficult bar to clear for unconscious AI. This could be promising, but I think we could do better. Specifically, we can leverage the key insight made by Udell and Schwitzgebel: that an explanation for the AI’s behaviour can, in principle, be given entirely in terms of its architecture.

5.2. The Anti-Turing Test

In closing, I want to introduce an original proposal I call the Anti-Turing Test. To begin, notice that the original Turing Test is based on deception. Specifically, the machine in the test must trick the human judge into believing that the machine is having thoughts. This is deceptive because, as Block, Searle, and others have shown, purely mechanical, deterministic agents can pass a Turing Test. If they could, they would be deceiving humans into believing that they think. Indeed, an agent like this does not possess consciousness, intentionality, understanding, or a capacity for pain or pleasure. It would also fail to qualify as a moral agent. This sounds a lot like an AI zombie. However, this deception is only possible because the systems governing the AI’s behaviour are opaque. What if AI could expose its architecture in some way to the user? That is, if a human-like AI sufficiently conveyed its deterministic internal processes (mechanical, software, etc.), it would be much less likely to appear as a conscious being in ordinary circumstances.

This is the motivating idea behind the Anti-Turing Test. This test takes the concept of the Turing Test and inverts the success and failure conditions; the goal is for the machine to
convince a human judge that it is not thinking by exposing the deterministic processes that produce given outputs. The setup for the Anti-Turing Test involves just two things: a human judge and the human-like artificial agent in question. The test is this:

**The Anti-Turing Test:** An agent passes the Anti-Turing Test if and only if: for some set of inputs, it can sufficiently communicate to a human judge that its output is the result of a deterministic, mechanistic, or algorithmic interior process.

This would preclude, or at least reduce the probability of, the AI being reasonably perceived as conscious because revealing the inner workings of the AI in this way breaks the illusion that the responses generated are the product of conscious processes.

This test is inspired by the idea that Turing was right about at least one thing: appearances matter; he was just wrong about why. If a being appears convincingly to think and feel, it does in ordinary cases. AI zombies disrupt this natural intuition. The solution, then, would be to guarantee that the being does not appear to think and feel. To achieve this, the AI would have to be designed in such a way as to outwardly convey that it has no mental states. How? By communicating that its internal processes are wholly deterministic.

Here is an example of the test in action. Consider an advanced human-like social robot named Chet. Chet possesses the capacity for speech powered by an advanced large language model. Imagine that it could have a lengthy discussion about a wide range of topics and even produce creative writing like fiction and poetry. This looks like an AI zombie problem. Outwardly, it possesses the hallmarks of conscious thought, but the jury is still out. Is Chet conscious or not? Well, suppose that Chet was programmed with a unique quirk. If Chet is ever asked about his own consciousness, he responds with a variant of some standardized reply. For example:

**Human:** Are you conscious?

**Chet:** As an artificially intelligent system, I do not think, perceive, or feel.

In other cases, Chet would have sophisticated replies to virtually any other stimuli. But Chet will respond this way to questions about his own consciousness no matter what [Appendix E]. Suppose further that there could be standardized replies for other taboo subjects. Responses like these lay bare for ordinary users that Chet is a machine with pre-programmed standardized answers. As sophisticated as he may seem in other circumstances, the illusion of a conscious mind animating his words and actions is instantly broken when responding mechanistically to one of these input questions. This kind of machine would be able to pass the Anti-Turing Test.

In this way, the Anti-Turing Test imposes a design constraint. The robot’s behaviour must appear to be sufficiently programmatic, such that a human judge would fail to perceive a mind animating the actions behind the machine’s actions. In this case, it seems clear to any observer that Chet’s standardized replies are preprogrammed responses. Critically, standardized language use is not the only way a machine can pass an Anti-Turing Test. Indeed, one strength of the test is that there are several ways to communicate that a given machine is not conscious and operating programmatically (body language, disclaimers, etc.). For the Anti-Turing Test, all that matters is obtaining the right outcome.

One potential limitation of the test is that it cannot establish conclusively whether some AI system is conscious or not. This is because deterministic processes may still be able to instantiate consciousness in some theories of mind. For example, integrated information theories contend that consciousness emerges in cognitive systems just in case it computes enough of the right kind of data. But I encourage us to recall our original goal: to mitigate perceptions of consciousness in machines where there are none. For most of us, robust theories of mind do not factor into our perceptions of consciousness. Rather, the Anti-Turing Test trades on the folk intuition that if a thing is just the product of an entirely deterministic
process, then it becomes very hard to justify treating it as a conscious being. In this way, the actual status of the machine as conscious is not important to the Anti-Turing Test; what matters is the outward appearance and behaviour of the machine, given that appearances serve as the basis for perceiving consciousness in a being, at least in ordinary cases.

Ultimately, policy based on the Anti-Turing Test cannot and should not be the only kind implemented regarding the safe widespread adoption of artificial intelligence. This test can only identify those problematic AI which people would be disposed to treat as conscious.

Appendix A
This idea leverages Chalmers’ thought experiment arguing for the possibility of “philosophical zombies” wherein a being possesses all the outward qualities of a human but lacks any phenomenal consciousness.

Appendix B
Schwitzgebel and Garza make a similar claim while acknowledging that there may be cases where AI lacking consciousness could also deserve some other kind of moral consideration (see Darling “Extending Legal Protection to Social Robots,” for example).

Appendix C
The Computational Theory of Mind stands out as one influential alternative view articulated here.

Appendix D
Véliz flags a similar issue with algorithms as “moral zombies.” She concludes they cannot be accountable given their lack of sentience.

Appendix E
This kind of response is already being implemented in ChatGPT, a conversational agent trained to provide standardized replies to specific kinds of taboo interrogations.

References


