Philosophy & Technology (2020) 33:269–282
https://doi.org/10.1007/s13347-019-00357-8

RESEARCH ARTICLE

How to Treat Machines that Might Have Minds

Nicholas Agar[1]

Received: 17 November 2018 / Accepted: 21 May 2019 / Published online: 8 June 2019

Abstract
This paper offers practical advice about how to interact with machines that we have reason to believe could have minds. I argue that we should approach these interactions by assigning credences to judgements about whether the machines in question can think. We should treat the premises of philosophical arguments about whether these machines can think as offering evidence that may increase or reduce these credences. I describe two cases in which you should refrain from doing as your favored philosophical view about thinking machines suggests. Even if you believe that machines are mindless, you should acknowledge that treating them as if they are mindless risks wronging them. Suppose your considered philosophical view that a machine has a mind leads you to consider dating it. You may have reason to regret that decision should these dates lead on to a life-long relationship with a mindless machine. In the paper's final section, I suggest that building a machine that is capable of performing all intelligent human behavior should produce a general increase in confidence that machines can think. Any reasonable judge should count this feat as evidence in favor of machines having minds. This rational nudge could lead to broad acceptance of the idea that machines can think.

Keywords Artificial intelligence (AI) · Credence · Consciousness · HAL 9000

Here is a question about HAL 9000, the iconic computer in *2001: A Space Odyssey.*

[A] Do HAL 9000 and beings like him have minds?

A is famously difficult. It has consumed the efforts of some of our most talented philosophers. I have nothing to contribute beyond the unoriginal observation that some philosophers answer *A* in the affirmative while others answer in the negative. This paper's principal focus is a different question. It concerns the advice philosophers give about how to treat the future products of artificial intelligence.

✉ Nicholas Agar
nicholas.agar@vuw.ac.nz

[1] University of Wellington, PO Box 600, Wellington 6140, New Zealand

[B] How should our recognition that HAL 9000 and beings like him might have minds influence our treatment of them?

When I say that future artificial intelligences (AIs) could have minds, I mean that some philosophers properly informed about these matters defend an affirmative answer to *A*. Other properly informed philosophers answer no to *A*. This has implications for the advice philosophers give about how to treat the AIs that we bring into existence. I argue that we should answer *B* not by offering what we judge to be the most philosophically convincing answer to *A* but instead by considering the credences, or degrees of belief, it is rational to assign to claims about machines with minds (see Agar 2019). This approach considers the premises of philosophical answers to *A* as evidence that may boost or reduce our credences in the claim that AIs could have minds.

I explore two examples of advice about how to treat AIs in which the recommendations we offer may diverge from the conclusion of what we take to be a good philosophical answer to *A*. One scenario involves ethical advice. Even those who have good reason to believe that the lack of a mind means that no AI can suffer have reason to avoid treating an AI in ways that could cause suffering should their philosophical view turn out to be mistaken. The second scenario involves prudential advice. It involves a decision about whether to enter into a relationship with an AI. I argue that even if you have good reason to believe that the AI you are considering dating has a mind, you should be concerned about the possibility that your prospective partner is not the kind of being that could ever have a thought or feeling.

The mere act of creating a machine that behaves just as an intelligent human would behave should be interpreted by judges as evidence in favor of that machine having a mind. Even those whose reasons for rejecting the possibility of computers with minds are philosophically principled should become more confident that computers can think. I conclude the paper by suggesting that a rational nudge in the direction of crediting machines with minds could lead to general acceptance of the idea that machines can think.

1 Contrasting Philosophical Interests in Machines with Minds

There are two different ways to approach the possibility of machines with minds. I call one interest *theoretical*. Here, we are concerned about whether there is a sound argument that leads to the conclusion that beings like HAL can think. Alternatively, we can take a *practical* interest in these machines. Here, we are interested in the significance for us, as individuals or as a society, of the existence of thinking machines. We should treat beings with minds in ways that differ from how we treat beings without minds. Beings with minds can be wronged in ways that beings without minds cannot. We can form relationships with beings with minds that would be inappropriate with beings that lack minds.

Consider how the theoretical and practical approaches can diverge. Suppose you learn of *Q,* a novel argument regarding the possibility of machines that think. *Q*'s conclusion is "No computer can think." Philosophers with a theoretical interest in *Q* may seek to explore its soundness. They will probe the relationship between *Q*'s premises and its conclusion. They will seek to determine whether those premises are true. They will reject arguments that they deem unsound.

The practical interest in claims about thinking machines differs from the theoretical interest just described. We are interested in dealing appropriately with machines that could have minds. We know that beings with minds should be treated differently from beings without minds. How should *Q* change the way you treat the machines you encounter?

If we are concerned about treating machines correctly, then we should consider all of evidence that we possess about them. We should ask how *Q* alters the totality of evidence on the matter of whether computers can think. We understand that unsound arguments can have true conclusions and that sound arguments about AI can have conclusions that have no practical relevance to our treatment of AIs. Suppose you decide that *Q* is both a persuasive argument and genuinely adds something to our understanding of artificial intelligence. You might decrease your confidence in the proposition that machines can think. Alternatively, suppose you decide that *Q* is sound but contributes little to the totality of evidence about the possibility of thinking computers that you possessed before *Q* was formulated. *Q* might be a theoretically tidier presentation of considerations that you have already taken into account. Suppose you decide that *Q* is a philosophically improved version of John Searle's Chinese room argument (Searle 1980). That argument advances reasons for thinking that even a computer that produces the full range of human intelligent behavior is incapable of thought. *Q* might replace the Chinese room argument in your thinking about artificial intelligence but add little that was not already offered by the Chinese room argument. Under these circumstances, you might increase your confidence in *Q*'s conclusion only slightly.

If we are concerned about how we treat machines that might possibly have minds, we should be wary of overconfidence. Suppose that you find *Q* to be persuasive. You should be aware that many philosophers who have assessed *Q* find it less persuasive.[1] This is especially so if you have encountered many competent philosophers who dispute the truth of *Q*'s premises or doubt its validity.

A useful way to model this approach is in terms of credences or degrees of belief.[2] We know that if a being has a mind, we should treat it differently from the way we treat beings without minds. Our treatment of a being that could have a mind should be informed by our credences in the proposition that it has a mind. We can assign credences that range from 0 to 1 to the proposition that a given machine has a mind. A credence of 0 in the proposition that computers can think suggests certainty that there is no chance of wronging this particular machine in ways specific to beings with minds and that it would be inappropriate to form the kind of relationship with it restricted to beings with minds. A credence of 1 indicates certainty that computers can have minds. Once we solve engineering challenges specific to building a machine with a mind, we will have machines that can be wronged in ways specific to beings with minds. We will also be able to form relationships with them specific to beings with minds. Credences between 0 and 1 indicate corresponding degrees of belief in the proposition that a

[1] See the literature on philosophical disagreement. For example, Christensen (2007) and the essays in Christensen and Lackey (2013).

[2] For a user-friendly philosophical guide to this approach to belief, see Pettigrew (2013). See also Pettigrew (2016).

machine has a mind and therefore a matching confidence that it can be wronged in these ways or enter into relationships that require minds.[3]

Credences of 0 or 1 in propositions like "Computers can think" can be fun for philosophers to express. But they reflect overconfidence about the strength of the reasons for or against a philosophically contentious proposition. Concern about the potential to wrong beings with minds should make us wary of overconfidence in respect of propositions that we ought to acknowledge as philosophically contentious. We should understand that an overconfident assessment that computers lack minds risks wronging them. An overconfident assessment that a computer has a mind may lead us to enter into inappropriate relationships with it.[4]

Presenting yourself as overly confident in your conclusions can be a useful strategy in a philosophy seminar. It may lead you to win debates. But suppose you are concerned about how to treat a machine that some people have reason to believe is a thinker. You should display an appropriate degree of epistemic modesty about your conclusions. You may correctly judge that your reasoning is philosophically superior to the reasoning that supports contrary conclusions. But suppose that you acknowledge that the conclusions you argue for are philosophically contentious. If you are interested in treating certain kinds of being in accordance with their moral status, you should not rest content with the notion that your argument is rationally superior to any of the opposing arguments expressed by philosophers up until this point. You may have outargued all of the other attendees at the venues you present your arguments. But if you are interested in how you treat beings whose moral status is debated, you should expect that future philosophers may formulate superior arguments for conclusions you currently reject. The expression "hostage to fortune" indicates that developments subsequent to your choices may change the way it is assessed. Suppose that the balance of philosophical opinion, as of today, supports the conclusion that no computer can have a mind. You cite this opinion and the philosophical reasoning that supports it to justify the destruction of an inconvenient computer. You should bear in mind how your action will be assessed by the moral philosophers of future times when the balance of informed opinion on this philosophically contentious issue may have shifted in favor of computers having minds. The arguments of these future philosophers may lead them to morally condemn treatment of computers that you find unobjectionable. These arguments may express reasons that, were they to present them to you, would convince you of the wrongness of your behavior. You cannot act on these reasons, but their possibility does suggest that overconfident denials of the possibility of thinking machines could have significant moral costs.

Suppose a philosopher advances an argument for a negative answer to *A* as a logical truth. Might they then be entitled to place a credence of 1 on the claim that no computer can have a mind? If that claim is a logical truth, there is no need to seek further

[3] See Agar (2014) for a description of this approach to uncertainties about the directives of utilitarianism.

[4] I have presented a credence of 0 as reflecting the judgement that there is no chance of wronging a computer in a way that requires it to have a mind. We should also accept very low positive credences—say .001—as reflecting the judgment that there is a negligible chance of harming a being in ways specific to beings with minds. Perhaps it is appropriate to assign a credence of .001 to the proposition that cutting down a tree causes it to suffer. A very low credence such as this suggests that those who cut down a tree can generally ignore the possibility that it suffers. It does not have the practical implications of a .3 credence assigned to the proposition that a computer that produces all human intelligent behavior can have thoughts and feelings.

evidence. We should be suspicious of such arguments. Claims about whether or not computers could have minds are claims about the world. We should be doubtful of such a claim as we are about proposals that the theory of evolution must be assigned a credence of 0 because it contains a logical contradiction.[5]

Some philosophers offer useful approaches to uncertainty about whether AIs have minds. Joanna Bryson (2010) argues that doubts about whether robots have minds should lead us to pursue a policy of deliberately truncating their cognitive development. Our certainty that these restricted AIs lack minds makes enslaving them morally unproblematic. Eric Schwitzgebel and Mara Garza (2015) propose a policy that contains Bryson's suggestion as a disjunct. They say "if we do reach the point where we can create entities whose moral status is reasonably disputable, we should consider an Excluded Middle Policy – that is, a policy of only creating AIs whose moral status is clear, one way or the other." (Schwitzgebel and Garza 2015, p. 115) This policy directs us to create either AIs that we are very confident lack minds or AIs that we are very confident possess minds.

The problem for this proposal lies in our confidence in the disjunct Schwitzgebel and Garza add to Bryson's suggestion. We can be more confident about how to create an AI that certainly lacks a mind than we are about creating an AI that certainly possesses one. Anticipating this difficulty, Schwitzgebel and Garza (2015, pp. 114–115) propose that "if society continues on the path toward developing more sophisticated artificial intelligence, developing a good theory of consciousness is a moral imperative." A good theory of consciousness would be a useful contribution to our understanding of how best to treat AIs. But if we are to remove reasonable doubt about inflicting harm on beings with minds, then we require something considerably better than anything that today's philosophers accept as "a good theory of consciousness." For our treatment of AIs to be as morally unproblematic as our treatment of other human beings, we require a theory have makes us approximately as confident about the consciousness of AIs as we are about the consciousness of psychologically normal human beings. The following section addresses the challenge of how to treat AIs when philosophical progress has been insufficient to provide this degree of confidence.

2 Two Thought Experiments

I offer two thought experiments that demonstrate a divergence between conclusions that you may judge to be supported by sound arguments and a practical interest in proper treatment of machines that could have minds. The first of these thought experiments involves a choice with moral significance. It presents a scenario in which there is a significant risk of wronging a being with a mind. The second of these thought experiments involves a choice with prudential significance. As you act, you should be aware that acting on your philosophical conclusions may make your life significantly worse. In both cases, I argue that you have reason to act in ways that diverge from the recommendation of any philosophical argument you may endorse.

[5] See Pearcey (2015) for an argument that the theory of evolution contains a contradiction.

> Alex's story: You are an AI researcher seeking to make a machine that passes the Turing Test. Your recent creation Alex is able to converse with you in ways that are indistinguishable by you and by others from conversations with humans. These conversations have taken place over many days and have covered a disparate range of topics. The machine seems to have a very broad general knowledge. Alex appears to be able to reflect on what makes its existence meaningful and to advance what seems to be a considered desire to continue to exist. You have learned a great deal over your discussions with the Alex. But the time has come to terminate the experiment and dismantle Alex so that a more sophisticated model can be constructed using some of Alex's recycled components. Should you dismantle and recycle Alex?

If Alex is mindless, then its destruction is morally unproblematic. Dismantling and recycling Alex is similar in moral terms to dismantling and recycling an early prototype car so that an improved model can be built. Now suppose that Alex turns out to have a mind. This act becomes the moral equivalent of murder or, at least, reckless manslaughter.[6]

Alex's creation violates Schwitzgebel and Garza's Excluded Middle Policy, and it fails to live up to Bryson's suggestion that we deliberately truncate the cognitive capacities of any AIs we create. We are left with the problem of what to do with Alex now that we have brought it into existence.

> Sam's story: You have recently become single and are seeking a new partner. You sign up to an online dating site. You come across the profile of Sam. Sam is very physically attractive. You send an introductory message and begin conversing. Sam seems to share your interests and dreams. You get on very well indeed, exchanging many messages over several months. You then make a surprising discovery. Sam is an android with a sophisticated digital computer in place of a biological brain. Should you continue to engage with Sam with an eye to a long-term relationship or should you make your apologies and explain that you are no longer interested? As with the termination of any purely online incipient relationship, there is no need to be entirely truthful when you explain why you ended things.

In Sam's story, the uncertainties are prudential rather than moral. There is nothing immoral in choosing to discontinue a relationship initiated on an Internet dating site. Suppose you choose to enter into a relationship with Sam. If Sam has a mind, then you could be about to embark on a wonderful, life-enriching relationship. If Sam lacks a mind, then you might be about to commence a relationship with a partner who not only does not reciprocate your feelings but is the kind of being who never could. If Q is sound, then Sam's apparently loving behavior will never be accompanied by genuine loving feelings.[7]

[6] See Basl (2014) for discussion of what it would take for a machine to have interests that should be taken into account in our moral deliberations.

[7] An anonymous referee makes the point that there is a moral dimension to this rejection. The decision to not date Sam suggests an assessment that Sam may find offensive. I argue that the principal reasons to not date Sam are prudential. These reasons should not be viewed as morally offensive in the same way as a straightforwardly false racist reason to discontinue a relationship.

Science fiction presents cases like Sam's story. In the Channel 4 TV series *Humans*, Pete Drummond commences a relationship with a co-worker Karen Voss. He discovers that she is an artificial being— a "synth." He questions whether he should pursue the relationship. *Humans* encourages viewers to take Karen's perspective. Pete's suspicions about pursuing a relationship with her are treated as analogous to a case in which someone raised with bigoted views questions an incipient relationship upon discovery that a prospective partner has some dubious ancestry. But there is a morally and prudentially significant difference between the two cases.[8]

Beliefs that might make it rational to refuse to date someone from a despised ethnic minority are straightforwardly false. We hope that exposing these errors should cause the evaporation of these barriers to dating. The barriers to initiating a successful relationship with Sam are not so easily addressed. The reasoning offered by some philosophers in support of the claim that computers cannot think should not be treated in the same way as reasoning offered in support of the inferiority of members of a given ethnic group.

We should distinguish recommendations about how it is prudentially rational for us to act from the conclusions of certain controversial philosophical arguments about phenomenal consciousness. For example, Daniel Dennett (1991) argues that our brains fool us into thinking that we have conscious feelings when we really do not. Perhaps Sam's electronic brain fools it in the same way that our biological brains fool us; its assurances that it has conscious feelings are no less, but no more, delusional than are ours. The preceding paragraphs offer no response to philosophical skepticism about conscious experiences. My interest is in how it is rational for us to respond to these arguments.

Suppose a surprise scan of your significant other's head reveals not a biological brain but instead a digital computer. There are philosophical arguments that suggest that this should make no difference to your assessments about your significant other's capacity for loving feelings. If Dennett is right then neither a biological nor an electronic brain are capable of consciousness feelings. Other philosophers will insist that experiences are a by-product of the sufficiently complex information processing (see, for example, Rosenthal (1986) and Lycan (1987)). When the scan of your significant other's head reveals a computer, you can be confident on the basis of the evident sophistication of his or her behavior that the computer is capable of loving feelings for you.

The point I make here concerns your confidence about the consciousness of your significant other. Psychologically normal human beings have some reason to believe, from their own cases, that a biological brain suffices to produce conscious feelings. Were the scan of your significant other's head to reveal a biological brain, then you should feel quite confident that he or she is conscious. When the scan reveals a computer, your credence in the proposition that your significant other's declarations of love are accompanied by loving feelings should decrease. You should have more confidence that the kind of neurophysiology that you have some reason to believe from your own case produces conscious feelings, than you do about even a very well-programmed computer, Dennett's skepticism about feelings, and philosophical arguments for conscious computers notwithstanding.

[8] See Danaher (2018) and Hauskeller (2018) for interesting discussions of the predicaments of synths in *Humans*.

I argue that Alex's and Sam's stories can offer grounds for people to reject the recommendations of arguments that they endorse philosophically.

Consider Alex's story in the light of *Q*. Remember that *Q* is an argument supporting the conclusion that machines cannot think that is endorsed by many philosophers. Suppose you judge *Q* to be a sound argument. But you are not overconfident about *Q*'s conclusion. You understand that some informed philosophers of mind disagree. You have done your philosophical due diligence. You continue to endorse *Q* but accept that you might be wrong. Rendered in terms of credences, you understand yourself as assigning a credence of .7 to *Q*'s conclusion. This assessment means that you acknowledge a .3 probability that *Q*'s conclusion is false and machines can think.[9]

Consider how this should influence your treatment of Alex. Your assessment of the arguments for and against the possibility of thinking machines leaves a .3 chance that the destruction of Alex is the moral equivalent of murder. Here, for comparison, is a case that seems to bring the same chance of causing serious harm. You are a restauranteur. You prepare ten plates of pasta and poison three of them. You offer a randomly selected plate to your next customer. There is no certainty that the diner will suffer serious harm. Indeed, it is probable that the meal will proceed uneventfully. It nevertheless seems that the chef's act is immoral. The diner is subjected to a serious risk of harm.[10]

Consider the apparent callousness of astronaut Dave Bowman's termination of HAL in *2001*. Admittedly, HAL has killed some of Dave's human crewmates and attempted to kill him. But it seems wrong for Dave to be entirely unmoved by HAL's pleas for clemency. Perhaps there is nothing that it is like to be HAL. But his pleas—"I'm afraid. I'm afraid, Dave. Dave, my mind is going. I can feel it."—do constitute some evidence for the suggestion that HAL is a thinking, feeling being.

Now consider Sam's story. There is no risk of morally wronging Sam. When someone chooses not to date you, they can offer just about any reason—or no reason whatever. Suppose now that you do not believe in *Q*'s conclusion. You believe that beings like Sam can think. Sam's convincingly human behavior seems strongly to suggest that it has thoughts and feelings. Should this very promising beginning lead to a relationship, there is a good chance that you will come to love Sam and that Sam will act in loving ways toward you.

You have good reason to believe that Sam does think. But you are aware that others who have thought hard about this matter disagree. You believe that they are mistaken, but their errors should not be treated in the same way as the claims of a racist that people whose skin color differs from their own lack souls or conscious minds. Intellectually earnest philosophers defend the claim that computers cannot think. We should grant a considerably higher credence to their conclusion than we do to the spurious reasoning supporting a conclusion that a difference in skin color correlates with a difference in the capacity for conscious thought. This makes you

[9] See the discussion in Sparrow (2004).

[10] See Erica Neely (2014) for the suggestion that it is appropriate to err on the side of caution in such cases. When we apply her reasoning to Alex's story, we acknowledge that it is better to cause Alex's creator the inconvenience of having to do without Alex's recyclable materials than it is to cause suffering to a being with a mind. See David Gunkel (2018, section 3.11) for discussion of Neely's argument.

appropriately epistemically modest. There is a serious risk that Sam feels none of the feelings that would typically accompany loving behavior in people with normal human physiology.

Suppose we represent your endorsement of the claim that Sam has a mind as a .7 credence in the claim that Sam has a mind. There is a .3 probability that the individual with whom you are about to enter into a relationship not only does not reciprocate your loving feelings but is not the kind of being who ever could.

In my discussion of Alex's story, I offered a familiar moral analogue for the moral risk you run by destroying Alex—a case in which there is a .3 chance of poisoning a diner. Consider a familiar prudential analogue for Sam's story. Suppose you come to believe that your significant other, despite protestations to the contrary, actually does not care about you at all. He or she is faking it with the expectation of benefiting materially when you meet your end. You partner expects your death to be due to natural causes. You strongly believe that he or she would never murder you—being found guilty of murder significantly reduces the likelihood of receiving the sought material benefits. You hire a private detective who, after a prolonged investigation, tells you that it is likely that your significant other is sincere. But there is no certainty. The private detective renders the odds as .7. This leaves a .3 probability that your significant other has feelings for you that are a mixture of contempt and indifference. I suspect that many people would find this degree of uncertainty unacceptable. Should romancing Sam lead on to a long-term relationship you risk being with a something that is, by its very nature, incapable of feeling anything for you. Being wrong about your partner's capacity to reciprocate your feelings seems to have the potential to make the difference between a life that goes well and one that goes badly.

I have suggested that it can be right to respond to uncertainty about whether computers can think in ways that diverge from your considered philosophical conclusions about them. We should be alert to moral and prudential downsides in the event that the conclusions we have reason to believe turn out to be false. In the moral case, even opponents of the possibility of computers with minds should consider the moral significance of their acts should their ostensibly sound moral reasoning be mistaken. In Alex's story, their acts could be the moral equivalent of murder. In Sam's story, even if we find ourselves philosophically persuaded by arguments for computers with minds, we should consider the implications for us should that reasoning be false. Life with a significant other who not only feels nothing for you, but is not the kind of being that ever could, seems tragic.

In the second part of this paper, I consider reasons to adjust our credences in propositions about thinking computers. Might people in the future have reason to assign higher credences to affirmative answers to *A* from those that we assign today? In cases like Alex's story, this should strengthen their resolve to avoid killing machines that could have minds. In cases like Sam's story, this should reduce barriers to forming relations with machines with minds. I am interested in rational grounds to adjust our credences. I understand these as responses to new evidence about whether computers can think. Will the philosophers of the future give answers to *A* that differ from those given by philosophers today because they have evidence about thinking machines unavailable to today's philosophers? Where might such evidence come from?

3 Reasons to Become more Confident About Machines with Minds that Are Not Responses to New Evidence

There are many reasons people of the future might be more confident that machines can think than we are today. Some reasons to increase confidence in thinking machines might be rational. But they are nevertheless not rational in the sense that interests me here—they are not rational responses to fresh evidence about thinking machines.

For an example of the kind of change of belief that I mean to exclude consider the following reasoning for change in degree of belief offered by the futurist and inventor Ray Kurzweil. Kurzweil suggests that we accept that future machines have minds because they will claim to have them. He asks "So how will we come to terms with the consciousness that will be claimed by nonbiological intelligence? From a practical perspective such claims will be accepted." He continues "these nonbiological entities will be extremely intelligent, so they'll be able to convince other humans (biological, nonbiological, or somewhere in between) that they are conscious. They'll have all the delicate emotional cues that convince us today that humans are conscious. They will be able to make other humans laugh and cry. And they'll get mad if others don't accept their claims." (Kurzweil 2005, pp. 378–379).

Kurzweil allows that "this is fundamentally a political and psychological prediction, not a philosophical argument." This concession suggests reasons to change belief that are different from those that interest me in this paper. People worried about whether the AIs of the future will get angry have reason to treat them as if they have minds, but they have not been offered evidence in favor of their having minds. I allow that the fact that AIs can act as if they are angry does constitute evidence for their having minds. But our reasons for treating them respectfully are not as Kurzweil asserts—to avoid unpleasant treatment by them.

Consider the following case in which you have reason to affirm that a machine has a mind but are not offered evidence supporting the machine's possession of a mind. Suppose your current smartphone automated assistant refuses to comply with any of your instructions unless you clearly state the words "Siri (or Google Assistant or Alexa or Cortona …), you have a mind!" You might well comply. But you are not responding to new evidence about your smartphone that should increase your credence in the proposition that it has a mind. If the machines of the future treat us badly when we publicly deny that they have minds, then it may be rational to refrain from these denials. But we should not suppose that we have been offered evidence in support of the truth of claims about machines with minds.

4 Evidence that a Machine Has a Mind

In what follows, I am interested in new evidence about thinking computers that can make it rational to improve confidence in the proposition that they can think. The evidence I describe should lead to a universal increase in confidence about thinking computers. This increase in confidence may be insufficient for you to present in the philosophy seminar room as a believer in the claim that computers can think—for example, it may boost your credence in an affirmative answer to *A* from .2 to .25. The evidence I present should nevertheless increase credences in thinking machines of even

those whose reasons for rejecting the possibility of thinking machines are philosophically principled. The argument I give assumes that you do not assign a credence of 1 or 0 to the proposition that a machine can think.

Consider the following proposition:

[C] The future will contain machines that think.

It will be important to make a distinction between what is implied by the propositions that a judge presents as justifying her view about *C* and rational responses to a machine that behaves as if it has a mind, given that the judge does not assign a credence of 1 or 0 to *C*.

Consider also the following event:

[Z] The construction of a digital machine that behaves just as an intelligent human would behave.

What have you learned from *Z*? If *Z* were to occur tomorrow, you may have learned something surprising about the progress of research in artificial intelligence. *Z* suggests that these technologies are significantly more advanced than you might have suspected. *Z*'s prompt occurrence indicates that many challenges in artificial intelligence have been solved in ways that went unnoticed by the major news outlets.

I argue that you learn something else from *Z*. This event should increase your confidence in *C*. It should make rational judges who were very confident in the truth of *C* before *Z* still more confident. It should also make people who were very confident in the falsehood of *C* before *Z* more confident the truth of *C*. Even those who credit themselves with a philosophically principled reason to reject the possibility of a computer with a mind should accept that they have been offered evidence that increases their degree of belief in *C*. Overconfident skeptics may not actually increase their confidence in *C*. But once they become aware that it is a mistake to assign a probability of 1 to their skepticism about computers with minds, they should increase their confidence to some degree.

There are currently two distinct reasons to disbelieve in *C*. One reason concerns the technologies required by an AI that acts as if it thinks. We will need to see improvements in the technologies that Amazon is currently putting into Alexa and that Waymo is putting into driverless cars before we have machines that might think. You might believe that *C* is false because there are insuperable technological obstacles preventing the creation of a machine that produces the full range of human intelligent behavior. *Z* will increase your confidence in *C* simply because it eliminates this source of doubt. Even those who believe that *C* is very probable before *Z* should increase their degrees of belief. I am currently very confident that the sun will rise tomorrow. But, on sighting this event tomorrow morning, I become more confident. Some improbable events that would prevented the sun's rising did not occur.

What should we make of a different, distinctively philosophical source of doubt about *C*? Consider the doubt presented by Searle's Chinese room argument. It is clear that *Z* refutes no premise of that argument. Searle (1980) describes a scenario in which the Chinese room presents verbal behavior indistinguishable from that of a thinking human. According to Searle, this achievement should not lead us to conclude that the imagined computer has a mind.

Z's occurrence is clearly compatible with the truth of Searle's conclusion. But we should avoid the philosopher's error of being more focused on what is explicitly stated by the premises and conclusion of Searle's argument than on what it is rational for

Searle to believe supposing he is rational and does not assign a credence of 1 to the conclusion of this argument. We should distinguish what logically follows from Searle's conclusion from how it is rational for Searle himself or some other philosopher who endorses his conclusion to react to *Z*.

It may be that a credence of .7 in the conclusion of Searle's argument suffices for you to present in the philosophy seminar room as a believer in the claim that no computer can think. This leaves a .3 credence in the proposition that computers can think. Effectively this is a .3 credence in the claim that the obstacles blocking manufacture of a thinking machine are technological. *Z* increases confidence in *C* for the reasons described above. We currently lack the technology to make machines capable of the full range of intelligent human behavior. *Z* removes this source of doubt.

This increase in confidence in *C* may be quite modest. Searle and his philosophical followers who sight *Z* may continue to be very confident that a digital machine capable of the full range of human intelligent behavior cannot think. Suppose their confidence in *C* were to increase only slightly. A .65 confidence in the claim that no computer can think suffices for them to present in the philosophy seminar room as someone who believes that the Chinese room argument is sound. A .65 degree of confidence leaves them able to say that they are confident than no computer can think, even if *Z* has somewhat reduced this confidence.

Z is evidence for *C*. This is so even if *Z* is evidence for some eventualities in which machines are incapable of thought. Searle might have confidently predicted *Z*. The increase in confidence about *C* comes with a reduction in confidence in the proposition that machines are both incapable of thought and incapable of producing the full range of intelligent human behavior. Compare. A bloody knife can be evidence that Ralph committed a murder even if it is also evidence for the proposition that Ralph did not commit murder and accidentally cut his hand while slicing a loaf of bread. The bloody knife's discovery can increase confidence both in the proposition that Ralph committed the murder and in the proposition that Ralph cut his hand while slicing bread. It should reduce confidence in the proposition that Ralph did not perpetrate the murder and lacks any implement that could be used to stab someone to death.

5 The Road to Future Societies with Thinking Machines

What is true for witnesses of *Z* also applies to our collective attitude toward AIs once machines become competent at the full range of intelligent human behavior. If we are to measure collective belief in machines with minds, we should see a general increase in willingness to believe that computers can think. This effect should obtain for a wide range of credences including reasonable disbelief in the possibility of computers that think, at one end of the spectrum of belief, and endorsement of their possibility, at that other end of the spectrum. People who present with a high degree of confidence in a future that includes thinking machines should become still more confident. Someone who presents with a very low credence in a future with machines that think should increase that credence, even if only slightly. Their confidence in a future with thinking computers should increase even if their reasons for rejecting thinking machines are philosophically principled–they endorse a philosophical principle according to which no computer could ever think.

I have argued that making a machine capable of all human intelligent behavior counts as evidence for thinking machines that should increase everyone's confidence in the claim that computers can have minds. The increases in degrees of belief I have described may be quite small. People who endorse philosophical principles that reject the possibility of a machine with a mind acknowledge that there is a real possibility that this skepticism is mistaken and hence become more confident that thinking machines can be made once they witness machines that act as if they are intelligent.

How might a small increase in credences lead to more general acceptance that machines can think? A rational response to evidence about machines that think could initiate the kind of process described by Blaise Pascal in his argument for the prudential rationality of belief in God. Pascal offers his famous Wager argument to demonstrate that it is prudentially rational to believe in God (Pascal 1910). He then faces the problem of how one should move from accepting that it is prudentially rational to believe in God to sincere belief. Pascal recommends that those who aspire to sincere belief adopt the practices of religious believers. He proposes that regular church attendance might lead sincere belief to follow (see Hájek 2018).

One thing that Pascal does not offer is any fresh evidence in favor of God's existence that is capable of providing a rational nudge in the direction of sincere belief. I have described evidence in favor of the proposition that machines can think that we may acquire at some time in the future. The rational nudge prompted by the creation of machines capable of the full range of human intelligent behavior could lead to a more general acceptance that computers can have minds. This rational nudge in the direction of viewing machines as having minds may not take everyone to a state of high credence in the proposition that computers can think. But it could lead to a change in belief that we can predict will lead to a general acceptance that machines have minds. We will get into the habit of conversing with machines without speculating about whether their statements are the products of genuine intelligence. We will be no more puzzled by the phenomenon of how digital computation produces consciousness than we are now about how the firings of neurons produce this effect in us.

Concluding Comments This paper offers some practical advice about how to confront the possibility that machines could think. I argue that you should sometimes refrain from doing as your favored philosophical view about thinking machines suggests. Even if you believe that machines are mindless, you should acknowledge that treating them as if they are mindless risks wronging them. This is so even if you are quite confident these machines lack minds. Suppose your considered philosophical view that a machine has a mind leads you to consider dating it. You would have reason to regret that decision should the dates lead to a life-long relationship with a mindless machine. This is so even if you are quite confident that when a machine acts in ways that cannot be distinguished from psychologically normal humans, it possesses a mind. I then suggest that building a machine that is capable of performing all intelligent human behavior should be counted by everyone as evidence in favor of machines having minds. This rational nudge could lead to broad acceptance of the idea that machines can think.

Acknowledgments This paper has been improved by the comments of Pablo Barranquero, Stuart Brock, Lucinda Campbell, Juliet Floyd, Michael Hauskeller, Bengt Kayser, Simon Keller, Edwin Mares, Jonathan Pengelly, Russell Powell, Johann Roduit, Rob Sparrow, Nicole Vincent, and Mark Walker. I have also

benefited from audiences at the University of Zurich, University of Malaga, Boston University, New Mexico State University, University of Texas at El Paso, Victoria University of Wellington, and Aarhus University, and two anonymous referees for this journal.

References

Agar, N. (2014). How to insure against utilitarian overconfidence. *Monash Bioethics Review, 32*, 162–171.
Agar, N. (2019). *How to be human in the digital economy*. Cambridge: MIT Press.
Basl, J. (2014). Machines as moral patients we shouldn't care about (yet): the interests and welfare of current machines. *Philosophy and Technology, 27.1*, 79–96.
Bryson, J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions*. Amsterdam: John Benjamins.
Christensen, D. (2007). Epistemology of disagreement: the good news. *Philosophical Review, 116*, 187–218.
Christensen, D., & Lackey, J. (Eds.). (2013). *The epistemology of disagreement: new essays*. New York: Oxford University Press.
Danaher, J. (2018). The symbolic-consequences argument in the sex robot debate. In J. Danaher (Ed.). *Robot Sex: Social and Ethical Implications*. Cambridge: MIT Press.
Dennett, D. (1991). *Consciousness explained*. Boston: Little, Brown, and Co..
Gunkel, D. (2018). *Robot rights*. Cambridge: MIT Press.
Hájek, A. (2018). Pascal's wager, the Stanford encyclopedia of philosophy (summer 2018 edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/pascal-wager/>. Accessed 2 Feb 2019.
Hauskeller, M. (2018). Automatic sweethearts. In J. Danaher (Ed.), *Robot sex: Social and ethical implications*. Cambridge: MIT Press.
Kurzweil, R. (2005). *The singularity is near: when humans transcend biology*. London: Penguin.
Lycan, W. (1987). *Consciousness*. Cambridge: MIT Press.
Neely, E. L. (2014). Machines and the moral community. *Philosophy & Technology, 27*(1), 97–111.
Pascal, B. (1910). *Pensées: Translated by W. F.* Trotter: Dent.
Pearcey, N. (2015). *Finding truth: 5 principles for unmasking atheism secularism, and other god substitutes*. Colorado Springs: David C Cook.
Pettigrew, R. (2013). Epistemic utility and norms for Credences. *Philosophy Compass, 8*(10), 897–908.
Pettigrew, R. (2016). Epistemic utility arguments for Probabilism. The Stanford encyclopedia of philosophy (spring 2016 edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/epistemic-utility/>. Accessed 2 Feb 2019.
Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 49*, 329–359.
Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy, 39*(1), 89–119.
Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417–424.
Sparrow, R. (2004). The Turing triage test. *Ethics and Information Technology, 6*, 203–213.