Outing A.I.: Beyond the Turing Test

The idea of measuring A.I. by its ability to "pass" as a human – dramatized in countless sci-fi films – is actually as old as modern A.I. research itself. It is traceable at least to 1950 when the British mathematician Alan Turing published "Computing Machinery and Intelligence," a paper in which he described what we now call the "Turing Test," and which he referred to as the "imitation game." There are different versions of the test, all of which are revealing as to why our approach to the culture and ethics of A.I. is what it is, for good and bad. For the most familiar version, a human interrogator asks questions of two hidden contestants, one a human and the other a computer. Turing suggests that if the interrogator usually cannot tell which is which, and if the computer can successfully pass as human, then can we not conclude, for practical purposes, that the computer is "intelligent"?

More people "know" Turing's foundational text than have actually read it. This is unfortunate because the text is marvelous, strange and surprising. Turing introduces his test as a variation on a popular parlor game in which two hidden contestants, a woman (player A) and a man (player B) try to convince a third that he or she is a woman by their written responses to leading questions. To win, one of the players must convincingly be who they really are, whereas the other must try to pass as another gender. Turing describes his own variation as one where "a computer takes the place of player A," and so a literal reading would suggest that in his version the computer is not just pretending to be a human, but pretending to be a woman. It must pass as a she.

Passing as a person comes down to what others see and interpret. Because everyone else is already willing to read others according to conventional cues (of race, sex, gender, species, etc.) the complicity between whoever (or whatever) is passing and those among which he or she or it performs is what allows passing to succeed. Whether or not an A.I. is trying to pass as a human or is merely in drag as a human is another matter. Is the ruse all just a game or, as for some people who are compelled to pass in their daily lives, an essential camouflage? Either way, "passing" may say more about the audience than about the performers.

That we would wish to define the very existence of A.I. in relation to its ability to mimic *how humans think that humans think* will be looked back upon as a weird sort of speciesism. The legacy of that conceit helped to steer some older A.I. research down disappointingly fruitless paths, hoping to recreate human minds from available parts. It just doesn't work that way. Contemporary A.I. research suggests instead that the threshold by which any particular arrangement of matter can be said to be "intelligent" doesn't have much to do with how it reflects humanness back at us. As

Stuart Russell and Peter Norvig (now director of research at Google) suggest in their essential A.I. textbook, biomorphic imitation is not how we design complex technology. Airplanes don't fly like birds fly, and we certainly don't try to trick birds into thinking that airplanes are birds in order to test whether those planes "really" are flying machines. Why do it for A.I. then? Today's serious A.I. research does not focus on the Turing Test as an objective criterion of success, and yet in our popular culture of A.I., the test's anthropocentrism holds such durable conceptual importance. Like the animals who talk like teenagers in a Disney movie, other minds are conceivable mostly by way of puerile ventriloquism.

Where is the real injury in this? If we want everyday A.I. to be congenial in a humane sort of way, so what? The answer is that we have much to gain from a more sincere and disenchanted relationship to synthetic intelligences, and much to lose by keeping illusions on life support. Some philosophers write about the possible ethical "rights" of A.I. as sentient entities, but that's not my point here. Rather, the truer perspective is also the better one for us as thinking technical creatures.

Musk, Gates and Hawking made headlines by speaking to the dangers that A.I. may pose. Their points are important, but I fear were largely misunderstood by many readers. Relying on efforts to program A.I. not to "harm humans" (inspired by Isaac Asimov's "three laws" of robotics from 1942) makes sense only when an A.I. knows what humans are and what harming them might mean. There are many ways that an A.I. might harm us that have nothing to do with its malevolence toward us, and chief among these is exactly following our well-meaning instructions to an idiotic and catastrophic extreme. Instead of mechanical failure or a transgression of moral code, the A.I. may pose an existential risk because it is both powerfully intelligent and disinterested in humans. To the extent that we recognize A.I. by its anthropomorphic qualities, or presume its preoccupation with us, we are vulnerable to those eventualities.

Whether or not "hard A.I." ever appears, the harm is also in the loss of all that we prevent ourselves from discovering and understanding when we insist on protecting beliefs we know to be false. In the 1950 essay, Turing offers several rebuttals to his speculative A.I., including a striking comparison with earlier objections to Copernican astronomy. Copernican traumas that abolish the false centrality and absolute specialness of human thought and species-being are priceless accomplishments. They allow for human culture based on how the world actually is more than on how it appears to us from our limited vantage point. Turing referred to these as "theological objections," but one could argue that the anthropomorphic precondition for A.I. is a "pre-Copernican" attitude as well, however secular it may appear. The advent of robust inhuman A.I. may let us achieve another disenchantment, one that should enable a more reality-based understanding of ourselves, our situation, and a fuller and more complex understanding of what "intelligence" is and is not. From there we can hopefully make our world with a greater confidence that our models are good approximations of what's out there.