A Test for Artificial Empathy

by

Adam Brandenburger* and Cheryl Loh**

Version 05/15/18

"Empathy, evidently, existed only within the human community"

-- Philip K. Dick, *Do Androids Dream of Electric Sheep*, 1968

1. Introduction

The recent rapid advances in AI have led a lot of people to ponder what a world of tomorrow populated with intelligent machines will look like. For some people, the promise is for a bright future for humanity where AI's help us in our homes, assist in caring for the elderly or raising children, and operate alongside us at work to boost our performance and even our creativity. An example is Kasparov [5, 2017], who describes the successes of human-machine partnerships in chess and extrapolates from these to a positive human future in the age of AI. Other people see the risk of an AI apocalypse where machines seize control of the planet and even threaten our physical existence. Such an existential catastrophe was described by Bostrom [1, 2014], and Stephen Hawking and Elon Musk have issued warnings in the popular media.

The dividing line between these two futures may depend on the issue of empathy, or, more precisely, on the possibility of artificial empathy (AE). If it is possible to build machines which reliably possess AE, then we can hope that their behavior will take into account human well-being as well as any objectives of their own. But if tomorrow's machines have their own goals and pursue them without regard to human welfare, then they may cause us harm --- even existential harm --- without caring. They will be sociopathic rather than empathic.

We do not pretend to have the blueprints on how to build AE into the machines of tomorrow. Rather, we propose a way to determine if AE has arrived. Our idea follows the tradition of the Turing Test and its descendants. Thus, we try to sidestep thorny definitional questions around the precise meaning of empathy and instead propose an operational test that compares human performance, in a setting which is generally agreed to require empathy, with the performance of machines (AI's) in the same setting. If the performances are not distinguishable, the AI passes the test.

Of course, we do not propose our test purely for intellectual reasons. We hope our test may also serve to suggest some useful directions for work, as in the case of other Turing-like tests.

2. The Turing Test

The Turing Test [12, 1950] was a brilliant move that avoided a lot of difficult questions raised by the question: Can machines think? Turing did not try to define the term "think." Instead, he proposed a test, based on observational equivalence, that such a machine would surely need to pass. We agree that humans think (whatever "think" means). So, we ask a person to interact remotely with both a machine and another person. The interaction should be of the type that is expected to trigger thinking on the part of another human. Turing chose a question-and-answer format, mediated by

* NYU Stern School of Business, New York, NY 10012, U.S.A.; NYU Shanghai, Shanghai 200122, China; adam.brandenburger@nyu.edu

** Conscious Consumer Lab, Shanghai 200031, China; cheryl@consciouslab.co

teleprinter. If the person cannot tell which of the two counterparts is the machine and which is the other person, we can conclude that the machine is thinking. More precisely, we can say that the machine is observationally equivalent to a thinking entity --- and we can reach this conclusion without danger of getting caught up in problematic definitional questions.

We want to mirror Turing's famous architecture as closely as possible, substituting the question "Can machines empathize?" for the original question "Can machines think?" To do this, we need to find a suitable analog to the question-and-answer exercise in the original test. This analogous exercise needs to be operationalizable in a form that the AI's of tomorrow (if not today) can follow.

3. Design Thinking

Our proposal for an exercise involving empathy which we hope could, not too far in the future, be undertaken by AI's as well as humans calls on an idea-generation technique called design thinking. The technique is widely used to help people come up with creative solutions to problems in the world of business and beyond (Brown [2, 2009]). Famous examples of outcomes of design thinking are the creation of Apple's first mouse and the development of talking heart defibrillators (that speak instructions during a medical emergency).

Here are the steps in the design-thinking process as specified by the leading teaching institution in this field (see Stanford d.school [9]):

- Empathize ("To create meaningful innovations, you need to know your users and care about their lives")
- Define ("Framing the right problem is the only way to create the right solution")
- Ideate ("It's not about coming up with the 'right' idea, it's about generating the broadest range of possibilities")
- Prototype ("Build to think and test to learn")
- Test ("Testing is an opportunity to learn about your solution and your user")

The process can cycle through these steps several times, as desired. In addition, each step is broken down into actionable components. A typical list of activities under the Empathize step might include: asking questions of current or intended users of the product in question; capturing via photo and video how users engage with the product in its current form; interviewing extreme users; and bodystorming via in situ role-playing and improvisation. (For definitions of "extreme user" and "bodystorming," see [10] and [11], respectively. This list of activities draws on [4].)

The next section will lay out how we envisage using design thinking within a Turing-like test for artificial empathy. A preliminary question, then, is if we could imagine AI's in some not-too-distant future being able to carry out a design-thinking exercise. For example, we might want an AI --- or a group of AI's --- to take on a challenge to design a better coffee cup according to certain criteria. We are going to assume that such an exercise will become possible at some point.

4. A Turing Test for Artificial Empathy

Many variants of the original Turing Test have been proposed over the years. The test for artificial empathy which we propose follows very closely the format of one such variant, called the Lovelace 2.0 Test (Riedl [6, 2014]). This test is designed to assess the ability of a machine to produce a creative artifact (such as a painting or story) that meets the standards of a human evaluator. In fact,

the author argues that certain kinds of creativity require human-level intelligence, and he therefore offers his test as a way of assessing both creative ability and intelligence. See [6] for more on this test, including a discussion of its potential advantages relative to earlier proposals (such as reducing the loophole where a machine is programmed to deceive evaluators in conversation).

The format of our test rests on [6], except that rather than being asked to produce a creative artifact, our machine is asked to tackle a design-thinking problem. We put forward and discuss some possible versions of our test.

Empathy Test Version 1: An AI is challenged as follows.

i. the AI must produce a solution *s* to a design-thinking problem *d*

ii. the solution *s* must conform to a set of criteria *C* where each $c \in C$ is a criterion expressible in natural language

iii. a human evaluator *h*, having chosen *d* and *C*, is satisfied that *s* is a valid solution to *d* and meets *C*

iv. a human referee *r* determines the combination of *d* and *C* to be not unrealistic for an average person

Empathy Test Version 2: Same as Version 1, except that the test is conducted with a group of AI's rather than with an individual AI.

Empathy Test Version 3: An AI and a person are challenged, in separate rooms, as follows.

i. the AI and the person must produce solutions *s* and *s′*, respectively, to a design-thinking problem *d*

ii. the solutions *s* and *s′* must conform to a set of criteria *C* where each $c \in C$ is a criterion expressible in natural language

iii. a human evaluator *h*, having chosen *d* and *C*, is satisfied that *s* and *s′* are both valid solutions to *d* and both meet *C*

iv. a human referee *r*, who does not see what is happening in the two rooms, is unable to tell which solution was produced by the AI and which by the person

Empathy Test Version 4: Same as Version 3, except that the test is conducted with a group of AI's (rather than with an individual AI) in the first room and a group of people (rather than an individual person) in the second room.

Riedl [6, 2014] explains that he specifies a set of criteria *C* in his test in order to make it Google-proof. The criteria serve to make the required artifact non-obvious, so that the machine cannot simply find an existing instance and output it. In our case, the specification of certain criteria is naturally part of a design-thinking exercise. (For example, in the better coffee cup exercise, one criterion might be that one can properly drink the foam as well as the liquid of a cappuccino without taking the lid off.) But the set *C* should also be chosen to include some surprising or even odd criteria, just as in [6], to make our test Google-proof.

Riedl [6, 2014] also explains that the requirement in his test that the criteria in *C* be expressed in natural language is not absolute, but the criteria must at least be equivalent to concepts expressible by a human mind.

An important question for our empathy test is whether it should be conducted on an individual AI or a group of AI's --- with the associated comparisons conducted on an individual person or a group of people. Moreover, a group could operate in an integrated manner or as a set of individuals working in parallel. Different design-thinking practitioners have their own preferred protocols. We prefer to be pragmatic in this regard, and so we describe the broad outlines of both an individual protocol (Versions 1 and 3) and a group protocol (Versions 2 and 4). We expect that technical feasibility will, in any case, be the main factor determining which protocols first become realizable.

Versions 3 and 4 differ from Versions 1 and 2 of our proposed test in adhering to the classic Turing format of pitting machine(s) against human(s) and seeing whether or not a referee can tell the difference. Versions 1 and 2 follow the format in Riedl [6, 2014] (and elsewhere), where a referee is asked to assess only the output of the machine(s) and say whether or not this output meets the human standard. Again, we have described the two approaches and leave this choice open, since we see no basis at present on which to favor one approach over the other in the case of our proposed empathy test.

5. Empathy or Theory of Mind?

A possible criticism of our proposed test is that it would really be testing for what in cognitive psychology is called Theory of Mind (ToM), not empathy. ToM is the ability to understand (at least, approximately) the mental states of others --- what others know, want, experience, feel, etc. But ToM is not feeling the mental states of others, which is what empathy involves. It is a cognitive not an affective process.

An AI that went through a design-thinking exercise, the criticism goes, would have worked through the steps involved, which include collecting various kinds of data from interviews, video, etc. of users. With the data, the AI may have arrived at a very good solution to the design-thinking exercise (such as a better coffee cup), which users greatly appreciate. The AI was effectively able to take on the perspective of users and solve their problem. But this sounds a lot like the acquisition on the part of the AI of ToM (or artificial ToM) about users. It does not sound much like empathy.

Actually, this point is already a criticism of the use of the term "empathy" in human design thinking, even before we get to the world of AI's. The first step may be labeled Empathize and the associated tagline may talk about caring (an affective state) about the lives of users. But there is no basis on which to assume that design thinkers actually feel, in addition to capturing and analyzing, the experiences of users. Perhaps, the whole association of empathy with design thinking is wrong (or, at least, not entailed). It is then a corollary that our proposed Turing test is not a test of empathy.

We think this line of argument is wrong-headed. But we put it forward in some detail because we think defusing it sheds some important light on what empathy operationalized really is about.

Turing [12, 1950] told us not to anthropomorphize intelligence. It would be beside the point, he said, to dress up a machine in artificial flesh, as part of testing its ability to think. Likewise, we should not insist that a machine be capable of feeling --- in some fashion akin to our human physiological sensations of feeling --- in order to say that it exhibits empathy. We should stick carefully with the operational approach, with its focus on behavior. In the case of artificial empathy (AE), the desired behavior is that an AI reliably and repeatedly acts in ways that take human interests

into account. (At the least, the AI does no harm. Better, it helps bring about desirable outcomes for us.)

With AE thought about this way, our design-thinking test looks conceptually appropriate. An AI (or group of AI's) will pass the test if it produces an outcome (such as the better coffee cup) that took into account and met the interests of the users in question. Of course, it is likely that the test should be repeated several times, with different design-thinking exercises, to check that one good outcome is not a coincidence and that human interests in the different exercises are reliably served.

Shanahan [7, 2015, pp.146-150] puts forward a very similar view on AE. He argues that the key to ensuring that a superintelligent AI does not take over the world, Machiavelli-style, is not whether it is really capable of feeling joy or sorrow for us, but how its reward function is designed. While we do not pretend to be able to offer a blueprint for building an AI which could pass our test, it does seem that the design of a good reward function for the AI will be very relevant.

6. Two Sides of Artificial Empathy

Our argument may come across as rather cold-hearted. It says that, in the end, AE actually should be understood more as ToM ('cold-heartedly' capturing and analyzing the relevant human experiences) combined with a suitable reward function for the AI. Such an AI only imitates empathy and, underneath, it functions more as a sociopathic mind-reader constrained by a reward function cleverly designed by its programmers. (See again Shanahan [7, 2015, pp.146-150], who goes on to express concern that it will be very difficult to design a reward function that does not produce destructive behavior in certain scenarios.) But imitation is, of course, at the heart of all Turing-like tests and, our argument goes, just because we associate human empathy with warm-hearted feelings, we must not insist that AE work the same way.

This said, there is a very important reason to think about the feelings that supposedly empathic AI's will invoke in people. A lot of effort around the world is currently going into building robots which, contrary to Turing's approach, are made to look human-like and which interact with us in a somewhat human-to-human manner. (Probably the two best-known current examples are Sophia from Hanson Robotics [3] and Pepper from SoftBank Robotics [8].) This is for very good reason. The bet is that, without these features (and a lot more such features), people's trust in AI's will be limited. We need to feel that AI's are not too different from us, in order to let them deeply into our lives.

In human relations, it is a good start when two people feel some commonality. We have a sense that the other person 'gets us.' But it soon becomes important to know if the other person also acts in ways that show real understanding of our needs. The evolution of AE may proceed similarly. Making AI's humanoid may get our relations with them off to a good start. But we will also want to be assured that AI's will reliably engage in behavior that solves problems for us and serves our interests. We propose our Turing test for AE as a yardstick for assessing when this second aspect of AE has arrived.

References

[1] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

[2] Tim Brown, *Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation*, HarperBusiness, 2009.

[3] Hanson Robotics, http://www.hansonrobotics.com/robot/sophia/.

[4] Interaction Design Foundation: "Design Thinking: Getting Started with Empathy," at https://www.interaction-design.org/literature/article/design-thinking-getting-started-with-empathy.

[5] Garry Kasparov, *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*, John Murray, 2017.

[6] Mark Riedl, "The Lovelace 2.0 Test of Artificial Creativity and Intelligence," December 2014, at https://arxiv.org/abs/1410.6142.

[7] Murray Shanahan, *The Technological Singularity*, The MIT Press, 2015.

[8] Softbank Robotics, https://www.ald.softbankrobotics.com/en/robots/pepper.

[9] Stanford d.school: "An Introduction to Design Thinking PROCESS GUIDE," at https://dschool-old.stanford.edu/sandbox/groups/designresources/wiki/36873/attachments/74b3d/ModeGuideBOOTCAMP2010L.pdf.

[10] Stanford d.school: "Extreme Users," at https://dschool-old.stanford.edu/wp-content/themes/dschool/method-cards/extreme-users.pdf.

[11] Stanford d.school: "Bodystorming," at https://dschool-old.stanford.edu/groups/k12/wiki/48c54/Bodystorming.html.

[12] Alan Turing, "Computing Machinery and Intelligence," *Mind*, 49, 1950, 433-460.