

# Ascribing Moral Value and the Embodied Turing Test

Neal Swisher<sup>1</sup>, Dobromir Dotov<sup>2</sup> and Anthony Chemero<sup>2</sup>

<sup>1</sup>Department of Instructional Technology, Drexel University

<sup>2</sup>Scientific and Philosophical Studies of Mind Program, Franklin & Marshall College  
tony.chemero@fandm.edu

## Abstract

What would it take for an artificial agent to be treated as having moral value? As a first step toward answering this question, we ask what it would take for an artificial agent to be capable of the sort of autonomous, adaptive social behavior that is characteristic of the animals that humans interact with. We propose that this sort of capacity is best measured by what we call the Embodied Turing Test. The Embodied Turing test is a test in which intelligence is operationally defined in terms of autonomous, adaptive interaction with the environment and with other animals. Three versions of the Embodied Turing test were performed with a SONY AIBO robot. Human participants were asked to differentiate between AIBO in a human-controlled mode and AIBO in a software-controlled mode. Our results indicate that the human participants were guessing at how AIBO was controlled. Our data reveals that people do not have enough experience with robots to accurately evaluate its behavior. This indicates that today's humans do not have enough experience with artificial agents to treat them as morally valuable.

## Introduction

In this paper, we turn the main question of the workshop—how would one build an artificial agent that behaves ethically?—on its head. Instead, we ask how one would build an artificial agent that deserves to be treated ethically, one that has moral worth? The two questions are, of course, not unrelated in that most things that treat others ethically are things that are seen as having moral worth. Note that the converse of this is not true: there are many agents that we regard as having moral worth but are incapable of ethical behavior. Infants and young children, non-human primates and pets are uncontroversial examples of non-ethical agents we take to be worthy of ethical treatment. Other animals are perhaps more controversial, but many would agree that all animals deserve at least to be treated non-cruelly. (This is why we bristle at factory farming and experiments by boys with reptiles and firecrackers.) Although it would surely be too strong to say that being worthy of ethical treatment is necessary for being capable of ethical treatment, it seems that there is a strong *de facto* connection.

So, what would it take to create an artificial agent with moral worth? It seems to us that a good way to start on this question is to ask what it would take for an artificial agent to be capable of the same sort of autonomous, adaptive, social behavior that is characteristic of the non-human

animals that humans interact with. To answer this second question, we have developed what we call the Embodied Turing Test. In what follows, we describe the Embodied Turing Test, and how it differs from Turing's original test. Then we present data from three experiments in which the Embodied Turing Test was performed. The results of these experiments are much more informative about the ability of human subjects to interact with artificial agents than they are about what (in terms of software and hardware) it takes to build intelligent, autonomous, social machines. They indicate that today's humans are simply unprepared to determine whether a machine is as intelligent, autonomous and socially capable as their pets, and hence as having similar moral value.

## The Embodied Turing Test

The Turing test is a standard test for the intelligence of a computer program (Turing 1950). In the test, a human subject has an instant-messenger-style conversation with two interlocutors, one of them a computer and one a human. The subject's job is to converse with the two interlocutors for a set period of time, attempting to determine which of the two is a computer and which is a human. A computer program that fools the subject some percentage of the time is taken to be intelligent. Turing himself suggested that any machine that fools human interlocutors 30% of the time would be intelligent. The Turing test, which no computer program has yet passed, provides an operational definition of intelligence: being intelligent is being able to pass the Turing test, i.e., having the conversational skills of an adult human.

The Turing Test has been criticized and defended many times over the years (e.g., Saygin, Cicekli, and Akman, 2000). Our main concern with Turing's original formulation is that it defines intelligence solely in terms of linguistic skills. This guarantees that all non-human animals are not intelligent. We think that it is simply obvious that many non-human animals are intelligent, in some sense of the word. We have devised a different form of the Turing test, which implies an operational definition of embodied intelligence, a kind of intelligence that humans and non-humans share. This sort of intelligence that is shared across species is manifest in autonomous, adaptive interactions with the environment and with other animals. The importance of this kind of intelligence has

long been recognized in the aLife and robotics communities (Brooks 1991, Clark 1997, Delancey 2003).

We used the capabilities of the Sony AIBO robotic dog to perform three different versions of a test analogous to the Turing Test, but that measures non-verbal, adaptive interactions. We call this test the Embodied Turing Test. In our test, the AIBO robot was operated in both autonomous and remote-control modes. In the former, it is controlled by software that we wrote; in the latter, it is controlled by a human in another room. In analogy with the original Turing Test, we suggest that a software controlled robot that routinely fools human interlocutors into believing that its behavior is under intelligent, human control possesses embodied intelligence. Our Embodied Turing Test, then, is a test of the intelligence that non-humans might possess, intelligence that is manifest in autonomy and the ability to interact non-verbally appropriately with the environment and other animals. This sort of embodied intelligence, it seems to us, is required for a machine to be deemed as possessing moral worth.

## Experiment 1

### Materials

The Sony AIBO robot used in these experiments was a model ERS-210A, equipped with an 802.11b wireless card. AIBO has a built-in memory-stick reader. Two different memory-sticks were used in this experiment. One contained our original software, which was used for the software-controlled mode. The other memory-stick used was the SONY Navigator software memory stick. This card was used for the human-controlled mode.

In the human-controlled mode, the human controller remotely controlled AIBO via a desktop PC in an adjacent room with the SONY Navigator software, using a wireless LAN router and AIBO's internal wireless card.

A USB joystick was used to control AIBO's movements, including walking, head movements, sitting, standing, lying down and kicking the ball. There were also on-screen controls for AIBO that included AIBO's vocalization (barking, growling, whimpering), and some pre-programmed movements such as the "shake paw" routine and the "good AIBO" routine, which caused AIBO to sit down, wag its tail and bark.

### Procedure

Participants were led into the experimental room by researcher #1. A second researcher would prepare AIBO before each subject, randomly choosing which means of control to use for AIBO (correcting the randomization every 10 subjects).

Researcher #1 would instruct the participant to sign a release form and then escort the participant into the arena.

After reading a briefing to the participant, researcher #1 would hand the participant AIBO's pink ball, wait for AIBO to boot up, and then inform the participant that he or she could begin. At the same time, the researcher would start a stopwatch and leave the room. Researcher #1 would wait in the control room where he could view a video feed from a tripod-mounted camera of the current participant's interaction with AIBO.

After five minutes, the first researcher would go back into the experimental room and inform the participant that the five minutes were over. The researcher would immediately walk over to AIBO and press the power button to shut it down. The researcher would then lead the participant to a table and administer Part I of the questionnaire. Upon completion, Part II of the questionnaire would be given. The first researcher would then administer the free response question, which asked the participant to elaborate on how they made their decision.

When AIBO was under software control, researcher #2 would watch a video feed of the experimental room while researcher #1 was briefing the participant. During AIBO's boot routine, researcher #2 would connect to AIBO via the wireless LAN in order to monitor AIBO's software. If a problem such as a software glitch was apparent, the experiment was stopped immediately.

When AIBO was under human control, researcher #2 would watch a video feed of the experimental room while researcher #1 was briefing the participant. During AIBO's boot routine, researcher #2 would put on the noise-canceling headphones and connect to AIBO via the wireless LAN using the Navigator software. Under these circumstances, researcher #2's visual and audio perception of what was going on in the experimental room was strictly limited to what was available to AIBO's sensors. That is, the only thing the experimenter could hear was the feed from AIBO's microphones, and the only visual link to the experimental room was via AIBO's video camera.

To be clear, there are two cameras involved in this experiment. One camera is AIBO's internal video camera. The other is a tripod-mounted video camera in the experimental room. When researcher #2 is controlling AIBO in the human-controlled mode, he is NOT using the tripod-mounted camera; he can only view the video feed from AIBO's internal camera.

### Results

Part I of the questionnaire consisted of one question, asking participants to rate how they thought AIBO was being controlled. The scale was from 1 to 5; 1 was "human," 3 was "don't know," and 5 was "software."

The mean score given by participants when AIBO was in the human-controlled trial was 2.65 (SD=1.23). The mean score given by participants when AIBO was in software-controlled mode was 3.10 (SD=1.52). Participants made the correct identification in the human condition if they scored AIBO as a "1" or a "2." Participants made the correct identification in the software condition if

they scored AIBO as a "4" or a "5." "Don't know," a score of 3, counted as failing to make the correct identification.

For 39 subjects, 20 (51.3%) made the correct identification and 19 (48.7%) failed to make the correct identification. This difference is not statistically significant.

The videos were coded for behavior of AIBO, and a significant difference was found in the total number of behaviors AIBO performed for each condition. AIBO performed significantly more behaviors (25.42) in the software-controlled mode than in the human-controlled mode (18.84),  $t=-3.9$ ,  $p<.01$ .

In software-controlled mode, participants made the correct identification 9 times (47.4%), and failed to make the correct identification 10 times (52.6%).

## Discussion

The difference in the scores for human- and software-controlled modes was not significant. Participants did not make the correct identification any more in the human-controlled mode than in the software-controlled mode.

In software-controlled mode, AIBO achieved a 52.6% success rate. That is, AIBO's software fooled participants into thinking that it was human-controlled 52.6% of the time. This far exceeds Turing's benchmark of 30%, so according to the original standard, AIBO "passed" the test and so has embodied intelligence. We do not draw this conclusion, however. There is no good reason to accept 30% as a benchmark; this an arbitrary target, one that Turing chose more as a prediction for future performance of computers than as an actual score to strive for.

Furthermore, because this was a one-trial design, subjects had no basis of comparison. AIBO's software fooled subjects into thinking that it was actually human-controlled roughly half of the time, but they also made the correct identification roughly half the time. Considering the data from the human condition and the overall test, it is likely that subjects were guessing at AIBO's mode of control.

To correct this problem, another experiment was conducted. We reasoned that if subjects could be exposed to both conditions, they would have more information on which to base their decision. They would not have to guess, and they would utilize less of their preconceived notions about computers and software.

Also, since AIBO performed significantly more behaviors in the software-controlled mode than in the human-controlled mode, we reasoned that AIBO might be too active in software mode, and that the software should be slowed down to match the human-controlled mode. For the second experiment, the software was completely rewritten in a more robust programming environment that gave us more freedom and access to more features of AIBO's hardware.

## Experiment 2

### Procedure

The procedure for the second experiment mirrored the procedure for the first experiment, with one crucial difference. Subjects were presented with both conditions: human- and software- controlled. The order was randomized: half of the subjects interacted with the human condition first, and half interacted with the software condition first. After interacting with AIBO for three minutes in each condition, subjects were asked to fill out Part I of the Questionnaire. Upon completion, Part II would be given.

### Results

The same questionnaires were used for experiment 2 as in the first experiment. The only difference is that all of the questions were asked twice—once for each condition.

The mean score given by participants when AIBO was in human-controlled mode was 2.81 and the mean score for the software-controlled mode was 3.25. For 32 subjects, 18 (56.3%) made the correct identification and 14 (42.8%) failed to make the correct identification.

The mean scores for human-controlled mode and software-controlled mode are not statistically different, and the frequency with which subject made the correct identification is not statistically different.

The most surprising data collected in this experiment was the difference in scores depending on which condition participants were exposed to first. The order of the conditions was randomized, with 17 of the subjects seeing software-controlled first, and 15 participants seeing human-controlled first. When shown the human-controlled condition first, subjects made the correct identification 11 out of 15 times (73.3%). When shown the software-controlled condition first, participants made the correct identification only 7 out of 17 times (41.2%). To put it another way, participants were very likely to guess "human-controlled" on the first trial regardless of the actual control-mode. 21 participants (65.6%) guessed "human-controlled" on the first trial.

### Discussion

The mean score for the human condition was a 2.81 and the mean score for the software condition was 3.25. Since the difference between these scores is not significant, subjects did not perceive a difference between the two modes of control. There was no change of participants' ability to guess AIBO's mode of control when the design was changed. Participants did just as poorly at identifying AIBO's mode of control in the one-trial design as in the two-trial. Being exposed to both conditions did not help the participants make their decision.

Turing's 1950 formulation of the test is generally taken to be the standard Turing test, although there is some controversy over his original intention (Sterrett 2000,

Traiger 2000). The 1950 formulation is a two-trial design. In 1952, Turing described a one-trial design of the test (Copeland 2000). It is unclear whether Turing thought there was a difference between the one- and two-trial designs because he describes different tests in different circumstances. Experiment 2 suggests that there is in fact no difference.

The fact that AIBO's software mode fooled subjects is surprising. It seems to us that the presence of a human controlling one of AIBO's modes should have been apparent. The human controller can respond to many more commands than AIBO's software understands. The human controller can also respond to much more complex and nuanced commands. The human controller is only limited in his control of AIBO by AIBO's physical limitations. For instance, the human controller understands all of the voice commands of the subject, but may not be able to follow through with the commands because of AIBO's limited repertoire of motion. The software, comparatively, understands very few commands, and not any complex commands of more than a few words.

Yet, the results from the second experiment confirm the results from the first experiment. Giving the participants some basis for comparison did not help them make the correct identification, and the software-controlled robot "passes" the test by Turing's criterion. Again, given the limitations of the robot's behavior under software control, we do not interpret this as showing that AIBO is intelligent. Instead, it seemed to us that subjects simply have too little conception of the capabilities of robots under differing modes of control. For example, one subject reported that the robot under human control was "too good" at tracking a ball, so it had to be controlled by software. This suggests to us the possibility that the question of whether the robot is human- or software-controlled confuses subjects.

### Experiment 3

#### Procedure

The procedure of the third experiment mirrored the second experiment with one aspect of the experiment changed. Subjects were shown both conditions (human- and software-controlled), but were not told of the presence of a human controller. They were told that they were to interact with two different computer programs controlling AIBO.

Subjects were told that they would be evaluating two different computer programs on the basis of which one was 'more intelligent.' Procedure was the same as in the second experiment, in which subjects were actually shown both conditions, in a random order. The questionnaires reflected the change in instruction; no mention was made of a human-controller. As far as the subject knew, both conditions were software-controlled.

#### Results

Ten subjects were tested with the revised procedure and questionnaires. 8 of the 10 subjects responded in the questionnaire that the condition that was actually human-controlled (unbeknownst to them) was more intelligent than the condition that was actually software-controlled. 2 subjects thought that the software-controlled robot was more intelligent.

#### Discussion

The third experiment reveals that once any mention of a human controlling AIBO is removed, subjects very easily make the 'correct' identification. That is, they say that the human-controlled mode is more intelligent than the software-controlled mode.

These results suggest that subjects *can* identify the presence of a human controlling AIBO and do see the human-controlled AIBO as more intelligent, but get confused when the question posed to them directly asks them to differentiate between a human and a piece of software.

#### Conclusion

Initially, our purpose in running these studies was to demonstrate a different kind of Turing test, one in which non-verbal intelligence is operationally defined in terms of appropriate interactions with the environment and an interlocutor. This test, we hoped, would do two things. First, we wanted to provide a definition of intelligence that could be applied to non-human animals and future intelligent animats. We did not expect that the software-controlled AIBO, which can respond to fewer than 20 verbal commands, would fool subjects. We did not expect, that is, to demonstrate that AIBO working under the software we wrote possessed embodied intelligence. Indeed, we do not want to have shown that, and insist that we have not. How, then, are we to interpret the results of our three studies?

We think that the best way to understand the results of these experiments is in terms of Joseph Weizenbaum's ELIZA software (1965). Weizenbaum's software responded to statements by a human interlocutor as would a Rogerian therapist, simply turning the human's statements into questions. Anyone today who plays with ELIZA is immediately struck by how unintelligent the software is. (ELIZA is built into the EMACS editor. Type `Escape-X doctor`, then `Return`.) Yet when Weizenbaum gave MIT students the opportunity to chat with ELIZA, many of them were fooled and had long conversations with the software, revealing details about their personal lives and, apparently, ignoring ELIZA's frequent non-sequiters and malapropisms. No one with the computer experience common to a teenager in today's Western society would be fooled. People were fooled by ELIZA in 1966 because at that point very few people had significant experience with computers, and hence very little basis for determining that

the computer program they were conversing with was not a human. This, though, was no reason to discredit the Turing Test as a measure of intelligence.

The present case is analogous to this one. Our software-controlled robot was able to fool subjects into thinking there was a human behind its actions not because it was intelligent, but because most people in 2005 have very limited experience with robots. Because of this very limited experience, our subjects had very little to draw on to establish expectations concerning the capabilities of a robot under human-control as opposed to under software-control. Our data indicates that they were simply guessing. And just as the plainly unintelligent ELIZA's successes is no reason to think that the Turing Test is invalid, we suggest that the successes of our manifestly unintelligent robot does not invalidate the Embodied Turing Test as a criterion for, and operational definition of, the non-verbal intelligence possessed by non-humans and robots of the not-too-distant future.

This has consequences for the main question of this workshop. In the Introduction, we pointed out that all currently existing agents that are capable of treating others morally are also taken as having moral value, though the converse is not true. This suggests to us that having moral worth is something like a prerequisite to behaving ethically. Our Embodied Turing Test indicates that humans are insufficiently experienced with robots to know whether to think of them as having the sort of intelligence characteristic of non-human animals. This suggests that contemporary, robotically-naïve humans also lack the capability to treat artificial agents as morally valuable. And given the propensity of contemporary humans to disregard the capacities of non-humans that allow them to feel pain, etc., we suspect that artificial agents will have capacities that make them morally worthy long before humans see them as such. We could, fifty years for now, have an animat rights movement.

## References

- Brooks, R. 1991. Intelligence Without Representation. *Artificial Intelligence*, 47, 139-159.
- Clark, A. 1997. *Being There*. Cambridge: MIT Press
- Copeland, J. 2000. The Turing Test. *Minds and Machines* 10 519-539.
- Delancey, C. 2003. Affect, Autonomy and AI. Paper presented at the Interactivism Summer Institute, Copenhagen, Denmark.
- Saygin, A. P., Cicekli, C., and Akman, V. 2000. Turing Test: 50 years later. *Minds and Machines* 10, 463-518.
- Sterrett, S. G. 2000. Turing's Two Tests of Intelligence. *Minds and Machines* 10.

Traiger, S. 2000. Making the Right Identification in the Turing Test. *Minds and Machines* 10, 561-572.

Turing, A. 1950. Computing Machinery and Intelligence." *Mind* 49, 433-460.

Weizenbaum, J. 1966. Eliza—A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* 9, 36-45.