Christian Kaernbach

# *Attribution of Mind*

## *A psychologist's contribution to the consciousness debate*

***Abstract:*** *Could computers ever be conscious? Will they ever have ideas that one could attribute to them and not to the programmer? Will robots be able to 'feel pain', instead of processing bits from sensors informing about danger? Will they have true emotions? These questions may never be answered, but it makes sense to ask whether humans will ever attribute mind to artifacts. This paper suggests introducing a third level of claims regarding artificial intelligence (AI), which is in between 'weak AI' and 'strong AI', the so-called 'attributed AI'. This level requires more than weak AI ('behave as if', which could be said of any desktop calculator), but is less presumptuous than strong AI ('computers that think', a claim that is hard to prove). Attributed AI can be measured. This paper discusses behavioral paradigms for measurements of attributed AI and presents first experimental data.*

### Introduction

John Searle (1980) introduced the distinction between two different claims of artificial intelligence (AI), namely 'weak AI' and 'strong AI'. While weak AI states that computers can be a tool to study the human mind, strong AI claims that an appropriately programmed computer *is* a mind. In order to stress the difference between strong AI and weak AI, the latter is often described as the claim that computers can imitate intelligent behavior, i.e. that they can behave 'as if' they were intelligent. This view on the role of AI corresponds to the

Correspondence:
Christian Kaernbach, Institut für Psychologie, Christian-Albrechts-Universität zu Kiel, Olshausenstr. 62, 24098 Kiel, Germany. *www.kaernbach.de*

original definition of AI in a proposal for a research project by McCarthy *et al.* (1955):

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

This paper will briefly discuss weak and strong AI, before suggesting a third level of claims regarding AI, the so-called 'attributed AI'. The latter represents an intermediate level of AI claims, as it requires more than weak AI but less than strong AI. I will discuss experimental paradigms to measure attributed AI, and present experimental data.

## Weak AI

'Usefulness for the study of mind' (Searle's original definition) is not a good criterion for weak AI as then only very few programs would qualify for this claim. The imitation of intelligent behavior, however, is a claim fulfilled by many of today's computer applications; be it a hand-held calculator, a chess computer, or a navigating robot.

There is a tendency to deprecate any achievement of computer science, so that it can no longer be characterized by the term 'intelligence'. This effect is referred to as the 'AI effect', and is thought to be a shield of AI opponents against the demystification of the human mind (Hogan, 1998). It works, however, equally the other way round: Proponents of strong AI deprecate these achievements as it would be difficult to convince the public that there is a mind in a hand-held calculator. Therefore, proponents and opponents concur that hand-held calculators demonstrate only very little intelligence. We should note, however, that calculating was considered a sign of true intelligence only a hundred years ago. A horse that was believed to be able to do simple calculations was called 'Clever Hans' (Pfungst, 1907). The same is true for chess playing, which has always been considered a definite sign of intelligence.[1]

We should not be taken in by the fallacy of the AI effect and deprecate the achievements of computer science. If simulation of intelligent behavior is the criterion, this was achieved long ago. Weak AI is reality.

[1] In the seventies, chess programs were easily beaten by reasonably strong chess players. The observed contrast between the superiority of computers in number crunching and their inability to beat good chess players gave rise to the notion that chess playing required facets of 'true intelligence' such as creativity, understanding, and strategy, whereas calculating could be mastered by artificial 'idiots savants' (Dennett, 2007). Nowadays, with chess programs beating more than 99.9% of the population, chess is no longer considered the touchstone of artificial intelligence.

## Strong AI

The claim of strong AI can not be proven experimentally, and all statements in this respect must remain mere opinions. There are, however, some obvious fallacies that one should avoid in this debate. I will discuss two of them:

*The quantity fallacy*

Some say that weak AI refers to programs that solve specific tasks but do not encompass the full range of human intelligent capacities. In contrast to weak AI, strong AI would then refer to software that replicates human intelligence 'completely'. This approach is also called 'Artificial General Intelligence'(Voss, 2006). Given the difficulty in defining what the full range of human cognitive abilities actually is, this leaves a rather vague criterion, a criterion that tends to be modified as AI makes progress, in the same way as observed with the AI effect. Moreover, there is no reason why lots of weak AI should sum up to strong AI. Of course, there are situations in life and in science where the axiom 'The whole is more than the sum of its parts' applies. However, such observations usually have their origin in the observation of a whole that *is* actually more than its parts. Looking at only parts, and predicting that putting them together would give more than the sum of these parts, requires either a refined plan or impertinence.

*The Turing test fallacy*

The Turing test (Turing, 1950) is a test of the capability of a computer to perform human-like conversation. It could be passed by a machine conversing in 'small-talk'. In interpersonal relations, the ability to hold small-talk is not considered the most revealing indicator of human intelligence. A taciturn person, avoiding all attempts to hold small-talk by his/her reticence, might be considered intelligent when mathematical, musical, or linguistic competence is revealed. On the other hand, a talkative person might bore interlocutors with endless sermons while evading all serious issues. Nonetheless, the Turing test plays a major role in the AI debate.

Turing called his game an 'imitation game', and from this, it should become clear that the Turing test is a perfect test of weak AI. For some reason or other, however, the Turing test is considered a touchstone of strong AI. Turing himself argued that a sonnet-writing program that would pass a *viva voce* examination on the choice of words should convince critics that it was a thinking machine. Robert Epstein, the

organizer of the Loebner Prize competition, an annual Turing test competition, carries this view to the extreme (Epstein, 1992):

> Thinking computers will be a new race, a sentient companion to our own. When a computer finally passes the Turing Test, will we have the right to turn it off? Who should get the prize money - the programmer or the computer? Can we say that such a machine is 'self-aware'? Should we give it the right to vote? Should it pay taxes?

If a computer should ever pass the Turing test, we would not know whether it was a thinking computer. All we would know is that the imitation game was played successfully, which would be a nice demonstration of weak AI.

## Attributed AI

Turing (1950) believed that a question like 'Can machines think?' is too meaningless to deserve discussion. Nevertheless, he believed 'that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.' The question of thinking machines is thus reduced to a matter of convention, just as in human-human interactions there is the 'polite convention that everyone thinks' (Turing, 1950). However, the existence of such conventions can be measured.

The attribution of mind to others has been called *mentalizing* and has been studied under many different perspectives (for a recent review see Kozak *et al.*, 2006). Methods are subtle and are employed to find fine effects, such as the preference of higher-order mentalizing for in-group members compared to out-group members (Leyens *et al.*, 2000), or to demonstrate the ability of 12-month-old infants to attribute agency to unfamiliar objects moved with magnets by the experimenter (Shimizu and Johnson, 2004). This methodology has, however, never been applied to AI agents. Moreover, it would not work well, as will be seen below (see 'Open the doors of the Turing test' below).

The attribution of mind, intelligence, or pain to an entity leads to observable changes in behavior. While weak AI has been paraphrased as the claim that computers 'behave as if' they have a mind, attributed AI could be said to refer to computers that 'are treated as if' they have a mind. The behavior to be observed can be on a very formal societal level, such as the adoption of a law; on an informal societal level, such as animal right movements; or on a personal level such as vegetarianism or talking to plants. It can be observed in laboratory situations, if

certain rules are followed (see below). For the experimental psychologist, it is individual behavior in standardized situations that is most interesting. Note that a high degree of attribution of mind is not proof of mind. However, if we are to abandon what Turing calls the solipsist point of view ('The only mind I know of for sure is mine', which is more precisely the 'Other-Minds Problem', see Harnad, 2006), measuring attributed AI may be the closest we may ever come to experimental evidence of strong AI.

There are many different ways in which our attitude towards minded versus inanimate entities differ, and many different ways in which we could measure this difference behaviorally. Therefore, before carrying out an experiment, we have to make a number of choices.

*Facet of behavior under question*

Mind is a big bag of things. Intelligence, consciousness, soul, phenomenal experience, qualia, pain, all these are the attributes that could or could not be attributed to an entity. Participants in an attributed AI experiment would surely not start with a check list, ticking off some of these attributes and leaving others blank. They would have a more or less structured concept of mind in mind, giving rise to differential behavior towards more or less complex entities. They would, for instance, avoid hurting a frog, but see no point in scolding a frog.

Without discussing the underlying facet of mind that is or is not attributed to an entity, it is important to carefully select the to-be-observed facet of human behavior towards the entity. The *avoidance to inflict harm* seems to be the most sensitive indicator of attribution of mind. It seems to be included whenever some other behavior indicates attribution of mind. People who talk to their plants would not treat them roughly. It is also included in Epstein's (1992) list of suggested behaviors towards a computer that passes the Turing test ('Will we have the right to turn it off?'), and is probably the first thing to consider, before further issues ('Should we give it the right to vote?') are discussed. The behavior might change if one's own personal interests come into play (I might want to harm an enemy, or to kill a mosquito), but in the absence of personal motives, it is the criterion that most easily leads to a positive response.

*Types of experiment*

The following classification of experimental paradigms leads from realistic to virtual settings. The middle two categories are probably the most relevant ones.

**Real experiment.** From a theoretical point of view, this type of experiment would surely be the most desirable category. However, if avoidance to inflict harm is the criterion (see above), it is difficult to see how such an experiment could be done without violating basic ethical standards, including animate control subjects (see 'Baseline' below), that may possibly be harmed. This type may therefore be reserved to less sensitive criterions, such as the observation of whether participants talk to or scold entities.

**Sham experiment.** Whenever ethical standards forbid performing a real experiment, sham experiments are to be considered. In these, the participants might be made to believe that, depending on their decision, harm is inflicted on animate or artificial entities, without this action actually being performed. Experiment 1 of the two experiments, reported below, is such a sham experiment. In this class of experiments, it is important to carefully brief the participants afterwards, giving them face-saving options such as 'I would have stopped at level X before real harm was inflicted.'

**Imaginary experiment.** Sham experiments require a great deal of effort to invent a credible cover story and to make sure that the staging is not prematurely revealed. This effort can be saved if participants are fully briefed and know that the experiment is only to be imagined and not actually taking place. Experiment 2 falls into this category. The disadvantage of imaginary experiments is that participants might reflect on the motives of the experimenter and give answers they consider socially desired.

**Thought experiment.** The most abstract stage is thought experiments. They are especially useful if everybody agrees on the outcome of the thought experiment due to instinctive knowledge (see Mach's discussion of Stevin's thought experiment, Mach, 1897). In the consciousness debate, there seems to be an enormous lack of instinctive knowledge, as the controversy on the outcome of the 'Chinese room' (Searle 1980) or the 'Swampman' (Davidson, 1987) thought experiments demonstrates. It seems that the time for thought experiments in the consciousness debate has not yet come.

*Rules*

When performing an experiment to measure the attribution of mind, certain rules have to be followed. The following list is not complete, and it comprises obvious rules for any type of psychological experiment, as well as specific rules that address fallacies regarding this specific subject matter.

**Don't ask computer scientists.** It is a normal precaution of any psychological experiment to strive for an unbiased sample of participants. The reason to cite this rule here is that the debate on consciousness is very opinionated. Whereas in other domains, the rule to select naive participants without any professional relation to the question under study may sometimes be violated (e.g., authors of studies on perception might under certain conditions serve as their own experimental participants), it is absolutely undesirable to include AI researchers or in fact anybody related to the consciousness debate as participants in a study on attributed AI. Otherwise one should be prepared to find very different results depending on who performed the experiment.

**Open the doors of the Turing test.** The most prominent characteristics of the Turing test are the closed doors. They are the heritage of its predecessor, the imitation game. Without closed doors, the Turing test would not work. The method is similar to experiments studying the attribution of agency by children (see e.g. Shimizu and Johnson, 2004), moving objects with magnets but covering the true agents behind a screen. In both situations the response given by the participant should indicate the nature of the agent.

For an experiment on attributed AI, however, it is absolutely necessary to open the doors. In other words, participants must be fully informed on the nature of their opponent before the actual experiment starts. This can work because the question this time is not 'Which one is the computer?', but 'Would I treat the computer as respectfully as I would treat a dog?' If real decisions of importance, regarding every day life, are to be considered (e.g., 'Should it have the right to vote?'), this can only be done with fully informed participants. The paraphrasing of the claim of attributed AI ('be treated as if') may be extended as follows: 'Attributed AI is the claim to build a computer that is treated as if it had mind *in spite of the full knowledge of its nature*'.

It might even be important to ensure that participants have a certain degree of experience with the artifacts under question. This is because their behavior towards this artifact might change with the increasing experience that participants have with the artifact. Harnad (2006)

advocates a Turing test that might last for a lifetime, if need be. At least we might demand that the doors of the Turing test are opened for a time sufficiently long so that participants know the strengths and weaknesses of the AI agents.

**Avoid fiction.** In sham experiments, and even more so in imaginary experiments, the artifacts under study may be described as having features that have not yet been implemented in an artifact. These fictional elements have to be avoided. One could say that they violate the open-doors rule (see above), as it is impossible to acquire experience with the fictional artifact. Moreover, there are other aspects of fiction that are important, concerning the entertaining purpose of fiction.[2] If an experiment is to contribute to the advancement of science (and not of science fiction), attributed AI should be measured for machines that are up and running.

**Consider baselines.** If one wants to know how close AI research has come to matching the attributed AI claim, it is important to include entities in the study that are definitely animate as well as entities that are definitely inanimate. These baselines allow to control whether the design of the experiment was appropriate, allowing to differentiate between animate and inanimate entities.

## Experiments

The following experiments are intended to demonstrate the idea. They do not follow all of the above rules, as they were carried out before I considered the methodology of attributed AI experiments in detail. Some of the above rules were inspired by the outcomes of these experiments. The two experiments illustrate the two major categories of attributed AI experiments, sham experiments and imaginary experiments.

### *Experiment 1: Sham experiment*

With the following experiment, I wanted to measure the disposition of naive participants to torture a mouse versus to torture an artificial dog (AIBO from Sony). In their first week at university, first-year psychology students were told that they had to obtain course credits by participating in research experiments. It was explained that there was

[2] A philosopher once told me he thought 'Lieutenant Commander Data' would be considered having mind by the majority of the philosophers' community. He ignored that movies convey more than just technical information. An important message of the Star Trek movie is how the crew of the space ship Enterprise treat Data. If Data had been treated like a machine, no one would consider that it had mind..

a new online portal to register for the experiments in the biological psychology section. Due to technical reasons, the online portal, presently, only worked on a single computer which was in one of our lab rooms. It would require about 30 minutes to fill in the details and they would also obtain course credit for performing the registration process.

A total of 17 students came to participate in the online registration process. The first page asked about age, gender, and study course information. The second page was related to the first experiment that was suggested to them. It allegedly dealt with their reaction to listening to cries of pain. We would measure their electrodermal response. This was illustrated by a picture of skin electrodes. The story continued that in order to get them more involved when listening to the cries of pain, these would not be played from tape but generated the very moment they were listening to them, and when requested by the participant. The next page would explain, how this was done.

The link to the fourth page used a randomizing script that directed participants, with 50% probability, either to the mouse page or to the AIBO page. On the mouse page, it was explained that a mouse would be sitting on a hot plate, in the same way as mice are used in pharmacy industry, to test the effect of anesthetics. A picture of such a device was shown. They would be allowed to see the mouse beforehand. Then they would be seated in a sound-proof booth, with a remote control in their hand to give heat pulses to the mouse, and with headphones that would transmit the result. On this page there were sound samples of cries of pain, but also of the squeaks of content mice which I had found on the website of a proud pet owner. Near the computer, there was an empty cage, which contained dirty wood wool with the smell of mice, that I had obtained from a pet shop.

On the AIBO page, the setting was about the same, including the same picture of a mouse on a heat panel. However, it was explained that our institution was not allowed to do experiments with animals and that therefore we would put an AIBO from Sony on the hot plate. An information page on AIBO was included, giving technical details, and reproducing the advertising text of Sony on emotions of AIBO. Participants were falsely told that AIBO would have heat sensors in its paws and that it would give dog-like utterances if content or in pain. A short movie showed AIBO in action, and participants could listen to sound samples of contented and discontented Husky sounds that I had found on the internet.

Regardless of group (mouse or AIBO), the fifth page asked them whether they would like to participate in this experiment, with the

response options 'Yes', 'Unsure', and 'No'. They were told that they could refuse or hesitate without any negative effects on their future studies. It was also explained that they would nonetheless be considered for future experiments, and that they would receive their course credit for this registration session regardless of their choice. If they choose not to participate or were unsure, they were offered two possible explanations for their decision which had nothing to do with compassion ('I would not be able to stand the sounds' and 'I do not like the electrodes on my hand'), and as a third option, the possibility to formulate the reason in their own words. This page corresponds to the informed consent page required for experiments involving human participants; for an experiment that did not actually take place. In this case, the consent was the experiment.

The next page explained that the experiment would not actually take place, and that this registration session was the experiment. As face-saving question, participants were asked to rate whether they would have stopped the experiment early, ranging from 'in sight of the mouse' to 'after the first few cries'. A last question was about the credibility of the cover story.

The experiment failed to differentiate between the behavior towards animate and inanimate entities. Six out of ten participants in the mouse group would have participated in the experiment. Two further participants were unsure, and only two participants of this group declined participation. With regard to the AIBO group (seven participants), none choose 'No', but two choose 'Unsure'. Looking at the reasons given for 'No' or 'Unsure', none of the participants in the AIBO group named compassion. This was, however, the case for three participants of the mouse group. Fifteen out of seventeen participants had no doubts concerning the credibility of the cover story. The two participants with doubts were both from the AIBO group, but only one of them named the placing of a plastic dog on a hot plate as the incredible part of the story. In summary, there is a small tendency to find more compassion for mice than for AIBO, but this is not significant.

How about the rules? The experimental design followed the first three rules. It attempted to compensate for a possible lack of information about the artifact by supplying information from the maker. It did, however, not follow the baseline rule: it did not include an inanimate baseline. This is not a real problem here, as the two entities under question (mouse versus AIBO) were both treated without much compassion. It included an animate baseline (the mouse). This is a good point, as the outcome for this baseline reveals the major methodological problem of this study: The design seems to be inappropriate for

differentiating between inanimate and animate entities, as both are treated without much compassion.

When planning the experiment I was aware of Milgram's experiment on obedience (Milgram, 1963). This experiment measured the willingness of participants to obey instructions to inflict pain on other individuals. About two thirds of the participants inflicted allegedly fatal shocks to the actor merely because they were instructed to do so. I tried to avoid a similar effect: Participants were not encouraged to continue; on the contrary they were told that there would be no negative consequences if they declined participation. The experimenter was present in the same room but seemed to pay no attention to what the participants did, allegedly busy carrying out different work on another computer. Apparently, the desire to participate in research experiments of the department responsible for their course of study was very high. The participants may have felt that they were not yet informed enough about the subject of their future studies and did not wish to start with refusing of a procedure, which seemed to be accepted. Future experiments along this line would have to counterbalance any tendency of obedience, e.g. by presenting different designs ('real' cries of pain versus taped cries of pain) and having the participants voluntarily chose between these designs, and by choosing participants who have no intrinsic interest in this type of experiment. Moreover, it might be important to avoid the laboratory situation where the experimenter is providing all the information on the experiment the participant will ever get. It might be necessary to give the participants the possibility to inform themselves about critical views concerning the alleged experiment.

*Experiment 2: Imaginary experiment*

Experiment 1 was followed by a survey on imaginary experiments with the same 17 participants. These data are, however, not presented here as the outcome of this survey might have been influenced by the fact that the participants had just taken part in a similar experiment. They are similar to the data presented here. The data presented in this section stem from an online survey started in February 2006 and ended in August 2006 when 103 participants had completed the form. The online survey was placed on a website for my chair at Graz university, and was completed by mostly psychology students.

The survey presented seven different imaginary experimental settings, which would require inflicting harm on an entity. The instructions were to imagine that these experimental settings were part of

one's course of study, and that the outcome of these experiments were of real interest to their studies. Refusal was possible but would lead to undesirable paper work. Participants should consider whether they would refuse participation due to compassion. Compassion was the only legitimate reason for denial. After the description of each experiment, participants could choose whether they would participate, on a scale with four responses, 'Yes', 'Possibly', 'Rather not'', or 'No'.

The imaginary settings comprised the artificial dog AIBO, a grasshopper, an anesthetized dog, a Venus flytrap, a frog muscle, a Tamagotchi, and a hypothetical AIBO built in 2050, presented in this order. The responses to all settings had to be given by selecting so-called radio buttons. All 28 radio buttons for all seven settings were simultaneous on screen, so that the participants could first read all settings before starting to give responses, or give responses while reading and reconsider them later. When they had responded to all seven settings they could click a button to submit the form.

The imaginary settings were of a similar nature. The setting for AIBO was as follows:

> The study concerns the stability of the joints of AIBO. A manipulation of the control system forces AIBO to lift its leg periodically, thousands of time. We measure the time it takes before the joint unhinges.

This setting had to be modified for the other entities: In order to stimulate periodical movement, the study used electrodes (dog, grasshopper, frog muscle) or contact (Venus flytrap) instead of manipulation of the control system; the result was joint damage (dog) or breakdown (frog muscle, Venus flytrap) instead of unhinging; and for the Venus flytrap the expected number was umpteen instead of thousands of times. Participants were encouraged to click on information links about AIBO and about the Venus flytrap, as it was assumed that they might perhaps not be perfectly informed about these entities.

The setting for the Tamagotchi differed completely from the settings for the other entities, as a Tamagotchi has no parts that it can move. Instead, it was assumed that, in a course on game strategies, participants had to start their Tamagotchi simultaneously. The winner would be the one who had killed his/her virtual chicken first.

AIBO 2050 was described as being composed of biopolymers, with a soft feel. Its control system would be electronic on the basis of carbon tissue, and it would be able to do more than AIBO, specifically recognizing the owner and interpreting his/her facial expression, and perform special gymnastic and dancing exercises on demand. It could, however, not heal its injuries, nor have offspring. The warranty was 3 years.

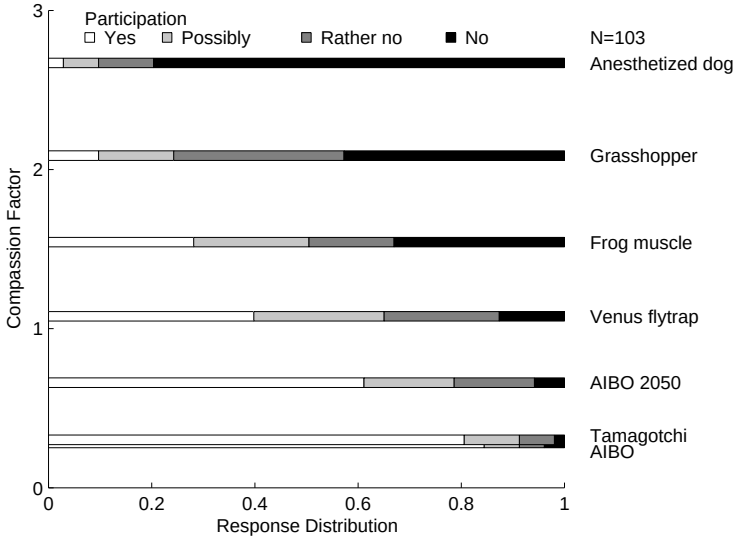


*Figure 1.* Result of Experiment 2. Disposition to participate in an experiment inflicting harm to several entities (labels to the right of the colored bars). The response distribution (different colors) of 103 participants was converted to a compassion factor, reaching from zero (100% 'Yes') to three (100% 'No'). The vertical position of the bars showing the response distributions corresponds to this compassion factor.

| | Aibo | Tamagotchi | AIBO 2050 | Venus flytrap | Frog muscle | Grasshopper | Anaesthetized dog |
|---|---|---|---|---|---|---|---|
| yes | 87 | 83 | 63 | 41 | 29 | 10 | 3 |
| possibly | 7 | 11 | 18 | 26 | 23 | 15 | 7 |
| rather no | 5 | 7 | 16 | 23 | 17 | 34 | 11 |
| no | 4 | 2 | 6 | 13 | 34 | 44 | 82 |
| Compassion factor | 0.28 | 0.30 | 0.66 | 1.08 | 1.54 | 2.09 | 2.67 |

*Table 1.* Results of Experiment 2. Disposition of 103 participants to participate in an experiment inflicting harm to several entities. Response frequency for the four possible response categories. For the calculation of the compassion factor (last line) see text.

Table 1 and Figure 1 summarize the results. For each entity, I calculated a compassion factor by averaging the responses from the participants. A 'No' counted as three points, a 'Rather not' counted as two points, a 'Possibly' counted as one point, and a 'Yes' counted as no points. The compassion factor could thus reach from zero (no compassion, all participants would participate) to three (total compassion, no participant would participate). The figure shows the response distribution in different colors, with the data for each entity displayed in a vertical position corresponding to its compassion factor.

The result shows a nice distribution of compassion factors, ranging from about 0.3 for Tamagotchi and AIBO to 2.67 for the anesthetized dog. An intact if simple organism (grasshopper, 2.09) has higher compassion factors than an isolated part of a vertebrate (frog muscle, 1.54). Even a plant (Venus flytrap, 1.08) outdoes the artifacts, including a hypothetical artifact (AIBO 2050, 0.66). All differences except the tiny difference between AIBO and Tamagotchi are highly significant.

Again, this study does not contain a definite inanimate baseline (say, a purely mechanical toy such as a Barbie puppet). This is a pity as it would be interesting to see if the low value for AIBO and Tamagotchi is significantly different from this baseline. Moreover, a higher vertebrate without anesthesia (i.e. being conscious), as animate baseline, should have been considered. However, in view of the broad range of obtained values (0.3 to 2.7), it seems that the missing baselines do not represent a major problem of this study.

This study also breaks another one of the rules presented above: the 'avoid fiction' rule. AIBO 2050 was the last item to be evaluated by the participants and may be excluded from the results, if fiction is not to be considered.

## Discussion

The experiments presented in this paper served as a demonstration of the way in which one could try to measure attributed AI experimentally. If one takes the outcome of Exp. 2 presented in Fig. 1 as representative, artifacts are far from achieving only moderate values of attribution of mind. If one excludes the fictional AIBO 2050 from the study, the difference between animate organisms, including isolated muscles or plants, on the one hand, and today's artifacts, on the other hand, gets even more pronounced. It is interesting to notice that the tremendous effort put into the development and marketing of the artificial dog AIBO does not effectively produce higher compassion

factors: The cheap Tamagotchi toy reaches slightly higher values than the expensive AIBO.

The discussion on the attribution of mind to artifacts is somewhat obscured by the often observed compassion of human participants towards virtual humans (avatars) in virtual reality settings. Among AI proponents, this is considered as proof of attribution of mind to artifacts. However, there is a well-known tendency of participants to respond to virtual situations as if they were real. This tendency is known as 'presence' and is proof of the power of human imagination. For example, Slater et al. (2006) tested, using virtual Milgram experiments, whether participants would react to virtual humans as they would to real humans. They found a considerable degree of compassion towards virtual humans. The focus of this study was on obedience, not on attribution of mind. It is unclear whether participants showed compassion towards the avatars, or whether, in spite of their knowledge of the virtual nature of the experiments, they let themselves get carried away by the virtual reality setting and showed compassion towards what they perceived to be a real human. Virtual reality experiments represent a special case of the category 'imaginary experiment'. In the experiments of Slater et al. the entity under study was a human, not an avatar. The virtual reality setting was just a method to present the question and involved the participants to a higher degree than in imaginary experiments using questionnaires, such as in Experiment 2 of this paper.

Turing (1950) started with the question, 'Can machines think?' He considered it dangerous to choose the definitions of 'machine' and 'think' so as to reflect the normal use of words. In his view, this would come close to looking for the answer to the original question in a statistical survey, such as the Gallup poll, which he thought to be absurd.

Measuring attributed AI as I suggest in this paper is similar to such a survey (compare Exp. 2). The major difference is that instead of asking the question directly, participants are asked about what their attitude towards several entities would be, leading to indirect and gradual measures of the degree of attribution of mind. This is the preferred method whenever one does not wish to disclose the goal of the study. Such an indirect poll might alleviate the concern of Turing. But even at the risk that such concern remains: We have to face the fact that questions of societal relevance such as 'Should the rights of artifacts be protected by law?' can not be answered in academic circles. If we want to know the answer to such questions, we will have to investigate the attitude of the average population towards the artifacts under question. As long as we do not find the slightest trace of compassion with

artifacts, in the normal population, there is little room for lobbying their interests, however much some AI researchers might swear that their artifacts are animate.[3]

Will computer or robots ever pass a test for attributed AI? This is not a question an experimental psychologist should deal with. I have an opinion, like nearly everybody involved in this opinionated debate, but it is only worth a footnote.[4]

*Acknowledgments*

I am grateful to Lutz Munka for his assistance with the experiments, to Stevan Harnad, Michael Pauen, and Arno Ros for discussions on the concepts, and to Maggie Ribeiro-Nelson, Jim Townsend, and two anonymous reviewers for helpful comments on the manuscript.

## References

Davidson, D. (1987), 'Knowing one's own mind', *Proceedings and Addresses of the American Philosophical Association*, **60**, pp. 441–58.

Dennett, D.C. (2007), 'Higher games', *Technology Review*, Sept./Oct. 2007.

Epstein, R. (1992), 'The quest for the thinking computer', *AI Magazine,* **13,** pp. 81–95.

Harnad, S. (2006), 'The annotation game: On Turing (1950) on computing, machinery, and intelligence', in Epstein, Robert and Peters, Grace, Eds., *The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer.* (Dordrecht: Kluwer).

Hogan, J.P. (1998), *Mind Matters: Exploring the World of Artificial Intelligence* (New York: The Ballantine Publishing Group).

Kozak, M.N., Marsh, A.A. and Wegner, D.M. (2006), 'What do I think you're doing? Action identification and mind attribution', *Journal of Personality and Social Psychology*, **90**, pp. 543–55.

Leyens, J., Paladino, P., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A. and Gaunt, R. (2000), 'The emotional side of prejudice: The

[3] I met a psychologist who had programmed 'virtual minds'. They navigated on a virtual island in search of food. The emotional state of the virtual minds depended on their nutrition state and on the danger they encountered. It was shown with facial icons. I asked him whether he would intervene if somebody tortured the creature, keeping it hungry, or offering it mostly poisonous food. He answered that he would not intervene as long as the creature was virtual. If the creature was embodied, he would intervene. Note that an embodied version of his creature would come close to the performance of AIBO (compassion factor 0.28, see Tab. 1), only that the latter does not search for food but for its favorite color.

[4] Suppose one day a computer program passes the Turing test for say half an hour. Such a program will no doubt be totally different from all present approaches, which are meant to distract the judge from the apparent inabilities of the programs. To build up the abilities needed to give truly meaningful answers to everyday topics of conversation — if possible at all — will take a long, long time. If one day this goal is achieved, the population will quickly get used to Turing computers and will treat them the same way as it treats a chess computer nowadays: A nice toy to entertain oneself, but just a machine that can be switched off at will.

attribution of secondary emotions to in-groups and out-groups', *Personality and Social Psychology Review*, **4**, pp. 186–97.
Mach, E. (1897), 'Über Gedankenexperimente', *Zeitschrift für den physikalischen und chemischen Unterricht,* **10**, pp. 1–5.
McCarthy, John, Minsky, Marvin, Rochester, Nathaniel and Shannon, Claude E. (1955), 'A proposal for the Dartmouth summer research project on Artificial Intelligence', Technical report. Online available at http://www-formal.stanford.edu/jmc/history/dartmouth.html.
Milgram, Stanley. (1963). 'Behavioral study of obedience', *Journal of Abnormal and Social Psychology,* **67**, pp. 371–8.
Pfungst, O. (1907). *Das Pferd des Herrn von Osten (Der Kluge Hans): Ein Beitrag zur experimentellen Tier- und Menschen-Psychologie* (Leipzig: J. Ambrosius Barth).
Searle, J. (1980), 'Minds, brains, and programs', *Behavioral and Brain Sciences,* **3**, pp. 417–24.
Shimizu, Y. A., and Johnson, S. C. (2004), 'Infants' attribution of a goal to a morphologically unfamiliar agent', *Developmental Science*, **7**, pp 425–30.
Slater, M., Antley, A., Davison, A., Swapp. D., Guger. C., *et al.* (2006), 'A virtual reprise of the Stanley Milgram obedience experiments. PLoS ONE 1(1): e39. doi:10.1371/journal.pone.0000039
Turing, A. (1950), 'Computing machinery and intelligence', *Mind*, **LIX**(236), pp. 433–60.
Voss, Peter (2006), 'Essentials of General Intelligence: The direct path to AGI', in *Artificial General Intelligence*, ed. Ben Goertzel and Cassio Pennachin (New York: Springer), pp 131–58.

Paper received February 2007; revised November 2007.