ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1670

The Post-Modern Homunculus

Eric Neufeld and Sonje Finnestad[1]

Abstract. Throughout the ages, magicians, scientists and charlatans have created life-*like* artifacts, some purported to be intelligent. In one famous case, *the Chess Player,* the intelligence was a little person hidden inside doing the thinking. Analogously, throughout the history of philosophy, and cognition, theories have arisen to explain intelligence in *humans*, but a philosophical problem with many such explanations is that they use what is called a *homunculus* argument – the explanation, upon scrutiny reveals a "little one" (homunculus) in the proposed mental apparatus that is responsible for thinking. For most of the era of computing, the Imitation Game, as so simply yet subtly put forward by Alan Turing, has been considered the gold standard for measuring this mysterious quantity, though recently Hector Levesque has pointedly argued that the time has come to abandon Turing's test for a better one of his own design, which he describes in a series of acclaimed papers. In particular, we argue that Levesque, who has cleverly found the 'homunculus' in the arguments of others, has essentially regressed the problem of intelligence to a homunculus in his own system.

1 INTRODUCTION

Understanding intelligence, artificial or otherwise, continues to challenge humankind. Homunculus arguments slip into many discussions, benignly or unintentionally, and it can take considerable scrutiny to find them. For instance, consider a theory of human vision that notes how the human eye works like a camera, with the lens projecting an upside-down image of the world on the retina, which an internal mechanism in the brain can watch and interpret. (This is a highly oversimplified presentation of the Cartesian Theatre [2].) Here, the homunculus is the aforementioned "internal mechanism" that has solved the problem of vision that the theory proposed to explain.

In a series of articles, several authors, in particular, Hector Levesque, revisit key foundational questions in artificial intelligence. Levesque, although he does not use the term *homunculus,* brilliantly uncovers one [4] in the celebrated Chinese Room thought experiment of John Searle.

Elsewhere [5,6], Levesque goes on to ask whether the Turing Test is obsolete and should be replaced by something else, and poses a very clever constructive alternative, a claim startling enough to make the *New Yorker*. However, a study of Levesque's papers suggests that, one, Levesque's new test also uses (inadvertently) a variation of a homunculus argument (although it was very hard to find), and two, that the Turing test is of a different character than other landmark tests in AI. In light of the special theme of this conference on *AI and Human Values*, we claim that Turing's test transcends formal philosophies of science, whereas other artificial intelligence tests are benchmarks. Judging humanity in any number of ways also transcends science, and so this makes a clear contribution to this central theme of human values.

The following elaborates these points.

2 LEVESQUE AND SEARLE'S CHINESE ROOM

Searle's Chinese Room argument [7] tries to separate outward behavior from true intelligence. "Imagine," he says, "a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program)." People outside the room pass in questions to the English speaker, who mechanically (!) consults his instruction book, retrieves certain answer symbols from the boxes, and returns them to the asker.

If the asker is pretty happy with the answers, the question "Is the behavior of the English speaker *intelligent*?" may be asked.

One camp argues that the English speaker's behavior is intelligent, and if we can build its computational equivalent, why do we care whether it understands what it is doing? This seems obvious. Another camp argues that the behaviour is *not* intelligent, because the English speaker has no idea as to what is being discussed, even if he or she is a fully capable human. This also seem obvious.

Both answers have merit. The first camp is concerned with pragmatics; the second demands a deeper notion of consciousness.

The question can also be asked in a different way. "Is the behaviour of the *system* intelligent?" Putting the question this way, the system includes the English speaker, the Chinese Room full of boxes, and the instruction manual.

Thirty years later, Searle [8] said "Computation is defined purely formally or syntactically, whereas minds have actual mental or semantic contents, and we cannot get from syntactical to the semantic just by having the syntactical operations and nothing else." In other words, "no". Searle's view is that neither the person nor the system is intelligent, because however clever its answers are the system, the system does not understand what it is doing.

Levesque [4] plainly notes, *there is no such instruction book* – mooting the discussion. He goes on to write that "Searle exploits the fact that we do not yet have a clear picture of what a real book for Chinese would have to be like." Why? Because such a book would have to have an answer for every possible question asked in every possible context.

That is, the *instruction book* is a homunculus!

For Searle's Chinese Room to function as described, some one (or thing) would have had to have already solved the problem of language understanding perfectly and incorporated it into an instruction book with a handy index.

[1] Department of Computer Science, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, email: eric.neufeld@usask.ca

Levesque takes a different tack from that described in the preceding paragraph, using an illustration he calls the Summation Room, which we do not reproduce for reasons of space.

3 WINOGRAD SCHEMAS AND THE TURING TEST

The genius of the Turing test (or *Imitation Game*) [9] is that it identifies intelligence with perhaps the most ordinary and basic of human activities – having a successful conversation in the estimation of a judge or jury. Turing thus proposed a way to test mechanical intelligence without actually defining it – a brilliant finesse.

Pointing to the Loebner competition, Levesque [5,6] states that the Turing Test "has a serious problem: it relies too much on deception". A program "will either have to be evasive … or manufacture some sort of false identity (and be prepared to lie convincingly)." "All other things being equal," says Levesque, "we should much prefer a test that did not depend on chicanery of this sort". "Is intelligence just a bag of tricks?" he asks. And so on.

To combat such "chicanery", Levesque proposes the Winograd Schema Challenge, which consists of multiple choice questions like this:

Question: The trophy would not fit in the brown suitcase because it was too *big* (*small*). What was too *big (small)?*

Answer 0: the trophy

Answer 1: the suitcase.

Levesque gives good arguments that tests can be constructed that are easy for humans to solve, yet are challenging for machines. Moreover, he provides a handy grading formula.

However, with the act of grading, the WSC becomes another benchmark, rather than a replacement for the Turing test. By 'benchmark', we intend many of the various challenges at which computers have already succeeded: tic-tac-toe, sliding tile puzzles, championship-level checkers, chess, or Go, Jeopardy, poker, and so on. The ability to objectively measure success in terms of win/loss, dollars, or speed allows us to define success in a manner that pays no attention to the machine's performance.

To understand this, consider the televised competition between Watson and two human Jeopardy champions. Watson almost enchanted until it was given, in the category *U.S. Cities*, the clue, "Its largest airport is named for a World War II hero, its second largest for a World War II battle." Watson answered "Toronto".

To the North American audience, this was a hysterical blooper, as most viewers knew Toronto to be in *Canada*, not the United States. To be fair, Watson knew it was guessing. Nonetheless, for many, this dispelled the illusion of intelligence, despite Watson's landslide victory, measured by money won.

The Turing test cannot use objective grading. It only requires that a human judge (or jury) be unable to distinguish the human's performance from the machine's. We believe that the existence of an objective grading scheme to decide intelligence in the WSC is equivalent to saying that a machine has been built that can decide whether another machine's performance on a multiple choice exam is intelligent or not.

There's no such machine! Unless, of course, someone has created one that can decide when another machine is intelligent.

It is interesting that Turing alternately suggested that judges or juries would decide the question of intelligence. There are at least two theories of why juries exist in the legal system [1]. One is that juries of peers tempered the decisions of judges, in the same sense the introduction of the House of Commons tempered decisions of the House of Lords.

The other theory, which [1] argues should be *central*, might be characterized as saying that the idea of justice ultimately resides in the minds of humans.

This brings us back to our earlier statement, where we stated that the Turing test transcends science. Let us be clear that we do not intend to enter the realm of the supernatural when we say this; it is only that in the Knowledge Representation community, it is appropriate to think of science in terms of the formal logical frameworks articulated, for example, by Kyburg [3]. But consensus on what science is has not been achieved there either – there are also frameworks going back to Popper and Quine, and many variations since.

4 CONCLUSIONS

The theme of this conference is *AI and Human Values*, and this work suggests that no computer has been able to impersonate a human and sustain the illusion for a reasonable length of time (despite some claims that the test has been passed), and that the ability to decide the success of such an impersonation remains a uniquely human task, even if computers are succeeding at various benchmarks perhaps sooner than expected.

This paper is a shortened version of [7]; this research was supported by the University of Saskatchewan.

REFERENCES

[1] R.P. Burns, The History and Theory of the American Jury. *California Law Review* 83(6):1477-1494, (1995).

[2] D. C. Dennett, "Darwin's dangerous idea." *The Sciences*, 35.3, 34-40, (1995).

[3] H.E. Kyburg, Jr., *Theory and Measurement*, Cambridge University Press (2009)

[4] H.J. Levesque, 'Is it Enough to get the Behaviour Right?', *Proceedings of IJCAI-09, Pasadena, CA*, Morgan Kaufmann, 1439:1444, (2009).

[5] H.J. Levesque, 'The Winograd Schema Challenge', *Logical Formalizations of Commonsense Reasoning, 2011 AAAI Spring Symposium, TR SS-11-06*, (2011).

[6] H.J. Levesque, 'On our best behaviour', *Artificial Intelligence* 212(1): 27-35, (2014).

[7] E. Neufeld and S. Finnestad, 'The Post-Modern Homunculus', *Proceedings of EGPAI,* to appear.

[8] J. Searle, 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3: 417–57, (1980).

[9] J. Searle, 'Why Dualism (and Materialism) Fail to Account for Consciousness', in R. E. Lee, ed., Questioning *Nineteenth Century Assumptions about Knowledge, III: Dualism.* NY: SUNY Press, 2010.

[10] A.M. Turing, "Computing machinery and intelligence." *Mind*, 433-460, (1950).