August Malueg

Response to "Truth in the Falsification of AI"

In their paper, "The Truth in the Falsification of Artificial Intelligence," author Mariah Jacobs argues that there are problems in applying the scientific methodology of falsifiability to the field of artificial intelligence. More specifically, they argue that Karl Popper's falsifiability is not possible within the field of AI because of our lack of understanding of the relevant concepts of consciousness or intelligence. That is, because we have no definitive or satisfactory understanding of consciousness or intelligent behavior, we are not able to hold the development or production of "strong AI" to the standard of falsifiability needed to confirm or disconfirm whether or not such development is valid science. Given this notion, they conclude that we must either 1. Aim to better define concepts in philosophy of mind such as "'personal identity,'" "'thinking,'" and "'intelligence'" so as to understand how they may be falsified, or 2. Adjust our scientific methodology to something "more appropriate" than Popper's falsificationism (Jacobs, 8).

Jacobs argues following the falsificationist scientific model that, "If a hypothesis cannot be disproven, or it is not understood what it would conceivably look like for it to be disproven, then there is no way to confirm or disconfirm with certainty whether it is valid science" (3). They go on to point out the various shortcomings and inherent limitations with behaviorist approaches to falsifying AI, such as the Turing Test, which rely upon the external or observable behavior of machines and the judgement of the human observer to determine whether or not these machines are in fact "strong AI." I would agree with Jacobs' assertions that behaviorist methods of analysis with regard to artificial intelligence ignore the interiority, "embodied perspective," or consciousness of whatever machine is being considered, and as a result, that we should not take

behaviorist approaches such as the Turing Test to be a sufficient way, in all cases, to determine whether or not a machine demonstrates the characteristics of "strong AI" (5) At the most, materialist approaches such as the Turing Test should be taken as indications of intelligence, and not an "adequate definition" of it (Jacobs, 5).

Following Jacobs, if we are to accept the implication of David Chalmers' 'philosophical zombie' thought experiment--that consciousness is something beyond the physical, and not physically or materially realized, then we are forced into quite a difficult position, for we are not able to analyze either the functions and observable behavior of machines, or their physical composition or structure, to sufficiently determine whether or not they demonstrate what is considered strong AI. If we are to embrace Chalmers' argument, and the arguments against behaviorist and functionalist approaches to falsification, it would seem that the falsificationist model of scientific analysis is altogether incompatible with regard to strong AI. Having thrown out all possibility of identifying human-level consciousness in machines with any material, observable behavior or characteristic, it follows that we are not able to hold the production of strong AI to *any* standard or hypothesis that could be falsified.

In light of this dilemma, and with no foreseeable solution to the "problem of other minds," or the notion that another human's consciousness is a non-verifiable possibility, perhaps a practical approach specifically to the production of artificial intelligence that reflects human intelligence is a materialist, behaviorist one. The problem that has been articulated in Jacobs' paper not only extends to artificial intelligence, but the human mind itself--the notion of solipsism is as much a relevant argument as Chalmers' philosophical zombie, for we cannot verify each others' conscious experiences much less those of machines we produce, or conclude that any materialist characteristics denote intelligence or conscious experience. Nonetheless, the

overwhelming feeling that other humans' behavior indicates consciousness proves enough for most of us to treat one another as conscious persons, and I would argue that, if our goal were to produce strong artificial intelligence defined as reflecting human-level intelligence, the final and most relevant approach is to look on a materialist basis for observable behavior that indicates human-like conscious experience. If we are looking to produce human-like intelligence, our falsifiable claim should be, 'this machine's behavior is like that of a conscious human.' From a pragmatic standpoint, it is precisely because we have no way of determining anything about the subjective or interior perspective of another being that we should revert back to the next-best option, and use behaviorist or other materialist analysis that takes physical indications of consciousness as its falsifiable hypothesis with regard to the production of human-like intelligence.

And yet, any behaviorist approach is inevitably too narrow to be a sufficient analysis for the development of all strong AI: because we are looking to signs of human-like behavior, this test will only be useful for the development of human-like intelligence, and will ultimately fail as a falsifiable claim when it comes to the production of other strong AI. It also does not grapple with the problem of embodied perspective. But without any certain knowledge of what non-human-like consciousness would look like, we should use a behaviorist or other materialist approach, while still hoping that developments in philosophy of mind and other fields make possible an understanding of how thinking and intelligence can be falsified from the standpoint of the subjective or embodied perspective.