Front. Philos. China 2016, 11(1): 3–20
DOI 10.3868/s030-005-016-0002-1

SPECIAL THEME

XU Yingjin

Does Wittgenstein Actually Undermine the Foundation of Artificial Intelligence?

Abstract Wittgenstein is widely viewed as a potential critic of a key philosophical assumption of the Strong Artificial Intelligence (AI) thesis, namely, that it is in principle possible to build a programmed machine which can achieve real intelligence. Stuart Shanker has provided the most systematic reconstruction of the Wittgensteinian argument against AI, building on Wittgenstein's own statements, the "rule-following" feature of language-games, and the putative alliance between AI and psychologism. This article will attempt to refute this reconstruction and its constituent arguments, thereby paving the way for a new and amicable rather than agonistic conception of the Wittgensteinian position on AI.

Keywords Strong Artificial Intelligence (AI), rule-following, psychologism, algorithm

1 Introduction

Artificial Intelligence (AI) is a branch of computer science which intends to realize or simulate human-level (or even supra-human level) intelligence via specifically programmed hardware. This ambitious pursuit naturally bears connection to the "computational theory of mind" (according to which the human mind is literally executing certain programs to exhibit intelligence), the plausibility of which is a major issue in the field of philosophy of mind. According to Searle (1980), there are basically two positions concerning this debate. One is the "Strong AI" thesis, namely that the human mind *is* a computer, and that a super-powerful computer, operating significantly beyond the capacities of present-day computers, could be equivalent to a human mind. The alternative to this is the "Weak AI" thesis, according to which a computer, however powerful, is at most something that barely simulates a human mind rather than

XU Yingjin (✉)
School of Philosophy, Fudan University, Shanghai 200235, China
E-mail: yjxu@fudan.edu.cn

something that is literally equivalent to it. Searle (1980) himself adheres to the "Weak AI" while other authors such as Penrose (1990; 1994) doubt whether even the "Weak AI" is tenable, to say nothing of the strong one. It is noteworthy that philosophical debates concerning the nature of both AI and the human mind have not merely been evoked by recent developments in computer science but rather have a long history even before the first generation of computers were put into service. For example, in Haugeland's (1985) reconstruction of early modern philosophy (when the possibility of mechanically executing human thoughts was already at issue), both Descartes (1637) and Leibniz (1714) can be viewed as critics of the "Strong AI" thesis, whereas Hobbes (1651) can be described as a prophet of the 20th century's so-called "symbolic approach to AI." Recently, interdisciplinary dialogues between philosophy and computer science (or cognitive science) have even involved contemporary continental thought, resulting in "Heideggerian cognitive science" and the like (cf. Kiverstein & Wheeler eds. 2012). Hence, it may be intriguing to think about what Ludwig Wittgenstein might say about AI, given his awareness of both the power and limitations of symbolic logic, as well as his personal connection to Alan Turing (in the 1930s at Cambridge), who is now widely identified as the founding father of computer science.

The consensus among Wittgenstein scholars such as Shanker (1998), Neumaier (1987), and Seidel (1991) is that Wittgenstein should be described as a foe of the "Strong AI" thesis. This view has also been affirmed by non-Wittgenstein scholars like Dreyfus (1992, 211), who has claimed that the ontological assumption of traditional AI is nothing but logical atomism, a position that Wittgenstein held in his *Tractatus* (1921) but that he abandoned in his *Philosophical Investigations* (1953, hereafter "*PI*"). Hence, it can, from this perspective, be easily deduced that Wittgenstein (as the author of *PI*) would not be at all sympathetic to traditional AI. Even Chalmers, whose own attitude towards Strong AI is more moderate, predicts that any *PI*-inspired scholar should challenge Strong AI by appealing to the "rule-following" feature of human behaviors that no computer program can mimic (Chalmers 1996, 329).

Among all the authors listed above, Stuart Shanker has provided the most systematic, and rather negative, evaluation of the Wittgensteinian position on AI. His 280-page book, *Wittgenstein's Remarks on the Foundation of AI* (1998), depicts modern AI as infected by both behaviorism and psychologism: two philosophically untenable positions which are supposed to have been greatly undermined by Wittgenstein himself. In this sense, Wittgenstein then undermined the very foundation of modern AI itself. Meanwhile, Shanker's only Wittgenstein-inspired proposal for doing AI and cognitive science, i.e. "to restore the focus onto an agent's social interactions, and away from that of a

self-modifying computer program" (Shanker 1998, xi), lacks "algorithmic" details. Hence, if Shanker's report were taken to be sound, it would be far-fetched even to describe Wittgenstein as a pioneer of some *non*-standard AI movement.

Thus, Shanker's work appears to be a significant obstacle to reconceiving the Wittgensteinian position on AI. To surmount it simply requires a careful examination of the most crucial arguments supporting his diagnosis.

2 Are Wittgenstein's Own Statements Really Hostile to AI?

Shanker's first major argument looks rather straightforward: since what we are discussing now is not whether *we* should be hostile to AI but whether *Wittgenstein* would be hostile to AI, *his* own statements should play a pivotal role in the debate. Since his statements are quite hostile to AI, it seems at first that Shanker's position is strong.

Still, Wittgenstein's two relevant statements bear consideration:

> **Statement No. 1** The problem here arises which could be expressed by the question: 'Is it possible for a machine to think?' (Whether the action of this machine can be described and predicted by the laws of physics, or, possibly, only by laws of a different kind applying to the behavior of organisms). And the trouble which is expressed in this question is not really that we don't yet know a machine which could do the job. The question is not analogous to that which someone might have asked a hundred years ago: 'Can a machine liquefy a gas?' The trouble is rather that the sentence, 'A machine thinks (perceives, wishes)' seems somehow nonsensical. It is as though we had asked 'Has the number 3 a colour?' (Wittgenstein 1935, 47, quoted in Shanker 1998, 2)

> **Statement No. 2** If one thinks of thought as something specifically human and organic, one is inclined to ask 'could there be a prosthetic apparatus for thinking, an inorganic substitute for thought?' But if thinking consists only in writing or speaking, why should not a machine do it? 'Yes, but the machine doesn't know anything.' Certainly it is senseless to talk of a prosthetic substitute for seeing and hearing. We do talk of artificial feet, but not of artificial pains in the foot.
>
> 'But could a machine think?'—Could it be in pain?—Here the important thing is what one means by something *being in pain* [or by *thinking*]. (Wittgenstein 1934, 105, quoted in Shanker 1998, 3)

Prima facie, both of Wittgenstein's statements appear to be opposed to the idea of "building thinking machines," and this may explain why they are so favored in Shanker's reconstruction of Wittgenstein's position. Shanker's reading of Statement No. 1 goes like this: Whether engineers can actually build a "thinking machine" has nothing to do with the truth of "A machine thinks," since this proposition is *a priori* false (just like "Number 3 has a colour" is *a priori* false) and hence cannot be verified by empirical findings. Statement No. 2 seems to stress this point from another angle by putting forward the following argument: If one were justified in attributing the predicate "artificial" to the noun "intelligence," one could also legitimately attribute the same predicate to the noun "pain," since both "intelligence" and "pain" designate mental entities. However, we all know that "artificial" is a modifier only applicable to physical entities, so anyone employing the term "artificial intelligence" or "artificial pain" would commit a "category mistake" by blurring the boundary between the mental and the physical.

If Shanker's interpretation of the two statements were accepted, one could immediately reach his conclusion that Wittgenstein does indeed undermine the foundation of AI by revealing the conceptual absurdities imbedded in it. Consequently, what Shanker's Wittgenstein is claiming here is unavoidably in conflict with two basic beliefs of AI scientists: (1) AI is by nature an empirical inquiry (cf. Newell & Simon 1976); and (2) "artificial" is definitely a legitimate attribute for "intelligence," since according to machine-functionalism (which is the tacit metaphysical assumption shared by the majority of AI scientists), "intelligence" is nothing but some functional property ultimately supervenient on certain physical configurations, and hence not as non-physical as it initially appears to be.

However, as I have claimed, the conflict between Wittgenstein and AI outlined above is created by Shanker's misunderstanding or over-interpretation of Wittgenstein's original statements. Let us first scrutinize Statement No. 1. It is noteworthy that the senselessness of talking about whether machines can think, which is undeniably pointed out by Wittgenstein himself, is itself conceptually based on his own idiosyncratic definition of "machine," according to which machines are artifacts whose behaviors can be predicted by the laws of physics (or by something else). That is to say, Wittgenstein's putative hostility towards AI is acceptable only when his definition of "machine" is assumed. (From now on this assumption will be labeled as the "Machine Definition Assumption," or MDA.) Or, to put it differently, Wittgenstein's argument is actually in conflict with contemporary AI only when MDA is actually applicable to AI.

Crucially, MDA is *not* applicable to AI. If MDA were true, then the behaviors of all contemporary AI devices should be predictable by laws, i.e., statements linking events that are supposed to be causally connected to one other.

Metaphysically speaking, if it truly exists, a law is destined to manifest itself when all intervening factors have been precluded. Hence, when some events are identified as causes inducing certain effects, the occurrence of certain events as causes should be predictable if the law involved both metaphysically holds and is epistemologically accessible. Let us apply this narrative to the functioning of an AI system. Since MDA has substantially involved the notion of "law," the behaviors of an AI system built in accordance with MDA should be predictable. That is to say, when the inputs of such a system are known to an observer, he, with the knowledge of the relevant laws, could in principle predict what outputs that machine would eventually deliver.

What are the instantiations of "law" in the context of AI systems? Conceivably, whatever they are, they should be applicable only to the local AI system rather than to the entire physical world, lest they exhaust the content of all existing physical laws and hence be only epistemologically accessible to an omniscient being à la Laplace's Demon. Any involvement of such omniscient beings will be distracting here, however, since their omniscience will immediately trivialize our focus on any local issue, including the possibility of Strong AI. Hence, if a more AI–relevant instantiation of "law" is expected to be realized locally enough, then it must only be applicable to the local system. Hence, there is only one category of qualified candidate left: programs describing the functioning of the local system, rather than physical laws describing the behaviors of the underlying physical substrate of the system. (Physical laws are prevalent throughout the entire universe, while algorithmic descriptions definitely concern the computing system itself.) Therefore, MDA can only be applied to an AI system when both the system's inputs and executing programs are epistemologically accessible to the predictor.

There are, however, some hard facts about contemporary AI which conflict with the last sentence of the last paragraph. One of the facts, among others, is that many AI systems' behaviors are unpredictable even though there is nothing unknown about their executing programs. Take connectionist systems (or artificial neural networks) as an example (cf. Garson 2010): a neural network consists of a large number of units, each simulating a neuron, joined together in a pattern of connections. Although the network works without processing a symbolic language, the programmer still needs to elaborate the network by preordaining how many layers of computing units should be involved, what algorithms are required for guiding the learning/backpropagation of the system, etc. Since the whole package of these parameters can be viewed as a program in a connectionist sense, the programmer does have the requisite law-like knowledge about the whole system. Even knowing that, however, he simply cannot predict what outputs such a system would deliver at a given time, since a connectionist system is "organic" and hence only roughly predictable in the long

run. Or, to put the matter more colloquially, even if he has confidence in the system's competence in completing a given task, he cannot tell precisely when and to what degree the system can do it. Meanwhile, he should also be prepared to reset the parameters mentioned above in case the system does not work as well as predicted. Thus, MDA cannot be applied to a typical connectionist system, and connectionist models, instantiating the currently most typical non-symbolic approach to AI, will not be troubled by Wittgenstein's negative comments concerning "thinking machines."

Some may raise the objection that even a connectionist system is "Turing-machine computable" in the sense that all its behaviors can be exhaustively simulated by a Turing-machine. Hence, if both the knowledge of the requisite machine-table (which is theoretically equivalent to a set of laws governing the behaviors of the entire system) and inputs fed into the system are available, then a precise prediction of the behaviors of the system will be easy to produce.

This objection neglects two further points, however. First, a reduction from the high-level description of a connectionist program to a specific Turing-machine table is merely theoretically possible, and neither feasible nor necessary in practice. Second, the predictability of the behaviors of the system is decidedly not the goal pursued by AI scientists, meaning that critics of AI seem to have greatly misunderstood them. In some sense, one can even say that unpredictability should be desirable in AI since it is routinely marked as an indicator of a machine's "being intelligent." (Consider how the unpredictability of responses could aid the designed system in misleading human judges during Turing tests). Or, to put it differently, once the dream of Strong AI comes true, such a system should deliver some *unpredictable* outputs due to its built-in agency, even though the programmer has knowledge about how this agency is realized by its functional configurations. Therefore, there is simply no way to convincingly apply Wittgenstein's philosophical comments on "thinking machines" to the intelligent machines that Strong AI advocates are taking pains to build.

Concerning Wittgenstein's Statement No. 2, wherein "thinking" and "pain" emerge as the most salient keywords rather than "machine": indeed, they are stressed by Wittgenstein himself in his last sentence: "'But could a machine think?'—Could it be in pain?—Here the important thing is what one means by something being *in pain* [or by *thinking*]." Obviously, this comment leaves open how to define "pain/thinking," and, consequently, how to answer the question "Could a machine think?" hinges on how to solve this semantic problem. The most natural interpretation of "thinking" or "pain," so to speak, is to identify it as some mental activity the nature of which is only epistemologically accessible from a first-person perspective. This reading immediately induces the following

argument supportive of Shanker's portrait of Wittgenstein:

(1) The very nature of thinking, or any other kind of mental activity, lies in its accessibility to the thinking agent from its own first personal perspective (a Cartesian assumption).

(2) To have pain is a typical type of "thinking" (*cogitare*) in this sense.

(3) Hence, a machine which has the capacity for thinking should have the capacity for having pain.

(4) However, we all know that machines can only simulate the behaviors of having pain rather than genuinely have pain.

(5) Hence, a machine cannot have access to any subjective feeling, if we take "to have pain" as the touchstone.

(6) Therefore, philosophically speaking, machines cannot think.

A Strong AI advocate who would like to refute this argument might focus on premises (1) and (4). Premise (4) looks most untenable since it has assumed a "neural-chauvinist" position, according to which phenomenal feelings of pain are only realizable in biological organisms, e.g. neural systems, and not in silicon-based substrates. Any "neuron-chauvinist" of this type, however, still owes us a further explanation for precluding the possibility of realized pain in silicon-based substrates. Premise (1) is also doubtable since it has assumed the absurdity of philosophical behaviorism, according to which the discourse concerning pain is linguistically reducible to the discourse exclusively concerning behaviors accompanying pain. However, as we all know, a genuine victory against behaviorism cannot be declared without fighting a single battle, and we simply cannot see where the requisite battle has taken place in premise (1). Moreover, since a Cartesian approach to mentality has been so saliently involved in premise (1), how such a premise could survive the currently widespread hostility against it is another problem. Therefore, the argument above either begs the question or requires further supplementation.

Advocates of Shanker's position may contend that even if the foregoing argument were not sound, it would not weaken Shanker's reconstruction, since his purpose is merely to reformulate what Wittgenstein intended to say, rather than to justify what he said. In other words, insofar as Wittgenstein's own comments are actually hostile to Strong AI, there is nothing significantly wrong with the picture that Shanker gives us.

However, there are two further reasons for casting doubt on the accuracy of his rendering of Wittgenstein.

First, as the last sentence of Statement No. 2 indicates, Wittgenstein is quite aware that his entire argument against the conceivability of a "thinking machine"

hinges on the meaning of "thinking" or "pain," yet he does not make any explicit claim to preclude a non-Cartesian reading of these terms. Or, to put it differently, if a reading of these terms in accordance with machine-functionalism (which is definitely non-Cartesian) is available, the implication of the whole passage will change dramatically. Obviously, Wittgenstein's own vagueness concerning this semantic problem leaves considerable room for conjecture about his original intention, and Shanker definitely underestimates the amount of this room.

Second, Statement No. 2 is taken from *Philosophical Grammar*, a text dating from an unstable period in Wittgenstein's intellectual development: the "Middle Wittgenstein," between his early stage and his later stage. More specifically, some of the ideas held by "Middle Wittgenstein" can be viewed as prototypes for the ideas of "Later Wittgenstein," whereas other ideas of this period would later be revised or even discarded. Thus, a more representative Wittgensteinian position concerning the conceivability of a thinking machine cannot be properly reconstructed if relevant comments in the later *PI* are neglected. Indeed, there are some comments in *PI* which conflict with the "Cartesian assumption" mentioned above, providing further reason to doubt the typicality of Shanker's portrait of Wittgenstein. To take *PI* §153 as an example:

> ***PI* §153** We are trying to get hold of the mental process of understanding which seems to be hidden behind those coarser and therefore more readily visible accompaniments. But we do not succeed; or, rather, it does not get as far as a real attempt.

This comment implies that the nature of "understanding," as a typical form of "thinking," does *not* lie in something "hidden behind" the linguistic outlook of "understanding," which is more visible and hence believed by some to be "coarser." This is tantamount to saying that there is no Cartesian picture of "understanding" hidden behind the portrait formed in the framework of language-games.

Moreover, Later Wittgenstein's hostility towards the Cartesian assumption is more typically revealed in his comments on "private language," wherein the Cartesian report of "*something* there all the same accompanying my cry of pain" is judged to be without any pragmatic force in language-games and hence lacking even minimal semantic content (*PI* §296). In other words, from his point of view, the very nature of pain does not lie in its being a "private object" privileged by a Cartesian subject, or, more metaphorically, "a beetle in the box" exclusively owned by its owner, but rather in its correlations with some visible behaviors gaining their face-values in intersubjective exchanges. In this process, the original "beetle" could even be "reduced" (*PI* §293).

Hence, put into a larger context consisting of both "Middle Wittgenstein" and

"Later Wittgenstein," Statement No. 2 does not appear to represent his mature attitude towards the relationship between "thinking" and "machine." The Wittgensteinian position on AI should hence be seriously re-evaluated.

Advocates of Shanker's position may contend that even if we focus on the Later Wittgenstein alone, his third-person approach to mentality by no means leads him to endorse the legitimacy of the enterprise of Strong AI, since there is something else in the Wittgensteinian language-games that no program can mimic. What is this "something"? Shanker's own reply is: rule-following.

3 Does "Following a Mechanical Rule" Actually Preclude "Mechanically Following a Rule"?

"Rule-following," as we know, is widely accepted as one of the most significant tenets of Later Wittgenstein. There are two more concrete ideas subsumed under this seemingly vague label: first, that "it is not possible to obey a rule 'privately'" (*PI* §202); and second, that whether or not an agent obeys a rule should be judged in a flexible manner, and cannot be judged either mechanically or overly arbitrarily. In Wittgenstein's own words:

> ***PI* §201** This was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule. The answer was: if everything can be made out to accord with the rule, then it can also be made out to conflict with it. And so there would be neither accord nor conflict here.

Since this passage may still seem ambiguous, a further explication is warranted. A helpful gloss is provided by Searle, wherein abstract discourses on "rule-following" are replaced by the example of "traffic-rule-following" (Searle 2004, 253–54). Suppose an agent is following the rule "Drive on the right-hand side of the road." The content of that rule must then be somehow causally related to his behavior, but this does not mean that the resulting behaviors of this agent should be entirely determined by the literal meaning of this rule, since there is still some room for flexibility, in which certain behaviors inconsistent with the rule, e.g., "Drive on the *left*-hand side of the road," are allowed if the derivation is needed for the agent to escape from some undesired situation, e.g., crashing into a huge rock lying on the right side of the road. However, this room for flexibility cannot be enlarged in an arbitrary manner, for otherwise any behavior literally disobeying the original rule would be judged to be obeying it. In other words, whether an agent is following a rule should be determined in a context-sensitive manner which cannot be mechanically construed.

Now we may link this observation to Shanker's picture of Wittgenstein. Although the "rule-following" discourse in *PI* is not directly relevant to AI, Shanker believes that Wittgenstein's philosophical wisdom can immediately lead us to the opposite side of the Strong AI thesis. Shanker's argument can be reconstructed as follows:

> (1) As we have seen above, to obey a rule (in the case of humans) by no means precludes the flexibility of revising it in a context-sensitive way, but rather assumes this flexibility. In Shanker's own words, "following a mechanical rule" is quite different from "mechanically following a rule," which leaves no room for even minimal flexibility. (Shanker 1998, 27)
>
> (2) It goes without saying that any genuine intelligent agent needs to "follow a mechanical rule" rather than "mechanically follow a rule."
>
> (3) However, a computer can only "mechanically follow a rule" rather than "follow a mechanical rule" flexibly. (Shanker 1998, 30–31)
>
> (4) Hence, a computer can never achieve genuine intelligence.
>
> (5) Therefore, the dream of Strong AI can never be realized.

The soundness of this argument heavily depends on the acceptability of premise (3). To demonstrate this, Shanker cites a comment made by computer scientist Donald Knuth:

> [A] favorite way to describe computer science is the study of *algorithms*. An algorithm is a precisely defined sequence of rules telling how to produce specified output information from given input information in a finite number of steps. [...] Perhaps the most significant discovery generated by the advent of computers [is that] an algorithmic point of view is a useful way to organize knowledge in general. (Knuth 1976, 38, quoted in Shanker 1998, 22)

Shanker believes that this testimony is clear enough to show that the "algorithmic point of view" favored by computer scientists paves the way to "mechanically following a rule" rather than "following a mechanical rule." More explicitly, since the execution of every single algorithm in any computer system should be precise enough for precluding any ambiguity, there is no way to algorithmically simulate the flexibility of human rule-following behaviors.

However, historically speaking, criticism of AI of this type is nothing but a reformulation of Descartes' criticism of the conceivability of intelligent machines in his *Discours de la Méthode* (1637), where the lack of flexibility and creativity of the outputs delivered by machines is taken to be supportive of a typically

Cartesian dualist point that the human spirit is irreducible to mere mechanical configurations. However, there is a rather questionable assumption imbedded in anti-Strong-AI arguments of this type, viz. that the algorithmic description of a computing system does preclude a non-mechanical level on a higher level. Let me reuse the example of connectionist models for explaining why this cannot be the case. As any AI textbook introducing connectionism demonstrates, a connectionist model specified for completing certain tasks, e.g., human-face recognition, is somehow flexible in contrast with the symbolic approach to AI. More concretely, if such a model has been fed with digital data from pictures of, say, Brad Pitt and George Clooney, and learned how to recognize both of the movie stars, then a new picture of Pitt, some features of which may be dissimilar from those of any other picture of Pitt ever fed into the system, will still be recognizable by the system. That is to say, some degree of novelty or fuzziness in the data can be perfectly tolerated by a connectionist system, and tolerance of this type is by no means "mechanical." If we remind ourselves that even a connectionist model is ultimately Turing-machine computable and hence describable from a purely algorithmic point of view, then we can immediately conclude that connectionism is a fairly convincing example of how lower-level mechanical executions can support emergent behaviors which are both non-mechanical and "rule-following"-like. Furthermore, one can easily imagine that there are infinitely many potential approaches to AI which share this emergent feature with connectionism. That is to say, counter-examples to Shanker's position abound.

Advocates of Shanker's position may contend that a connectionist system does lack a pivotal feature possessed by a genuinely "rule-following" agent, viz. the intellectual capacity for justifying his rule-following behavior, e.g., justifications such as, "Since I want to be a good citizen, I will always stick to the local traffic rules." But this rejoinder cannot work either. On the one hand, demonstration of "justification" is not entirely infeasible for a computer system, if "justification" is construed as a meta-description of the system's algorithmic executions. (Meta-description of this type is usually dispensable for the sake of efficiency.) On the other hand, even for a human agent it is not always possible to explain why a certain behavior is undertaken, a point stressed by Wittgenstein himself when he says that "Explanations come to an end somewhere" (*PI* §1). In this sense, the possession of some unlimited capacity for justifying given behaviors is too onerous a demand of humans or of machines. It is simply unfair to ask machines alone to fulfill this request.

The upshot of this section is that the "rule-following" feature of language-games has nothing to do with the plausibility of the Strong AI thesis. The former cannot be used to undermine the foundation of the latter.

4 Does Modern AI Assume a "Psychologistic" Position Incompatible with Wittgenstein's Philosophy?

"Psychologism" is a label historically attached to the position that the normativity of logical rules can be systematically reduced to features of findings via psychological inquiries, whether they are conducted in an empirical or transcendental manner. Leading figures in this intellectual movement include Friedrich Eduard Beneke (1798–1854), William Hamilton (1788–1856), John Stuart Mill (1806–73), and Rudolf Hermann Lotze (1817–81). As readers familiar with the historical background of both phenomenology and the early stages of analytical philosophy may know, a global battle against psychologism figures significantly in the agendas of Husserl and Frege, who were both eager to defend the irreducibility of logical norms. However, since there are virtually no prominent "psychologistic" thinkers in the traditional sense nowadays, many, including the author of this article, believe that topics related to psychologism would only interest historians of philosophy. Shanker however has a different view of this issue. He argues that psychologism is still alive in its updated form, i.e., "cognitive psychologism," which revives some obsolete ideas from classical psychologism in a brand new framework provided by the "cognitive turn" in 1960s. Shanker even believes that this problematic position is integrated into the philosophical foundation of AI, and that a Wittgenstein-inspired criticism of AI will therefore be more convincing if AI's psychologistic element is precisely targeted in advance.

The success of this strategy for making Wittgenstein into a foe of AI very much depends on whether or not there is actually a link connecting psychologism with current conceptions of AI. The question becomes: where can we find evidence for such a link? Its existence, Shanker believes, is confirmed by the AI scientists Newell and Simon, whom he quotes:

> We may then conceive of an intelligent program that manipulates symbols in the same way that our subject does—by taking as inputs the symbolic logic expressions, and producing as outputs a sequence of rule applications that coincides with the subject's. If we observed this program in operation, it would be considering various rules and evaluating various expressions, the same sorts of things we see expressed in the protocol of the subject. If the fit of such a program were close enough to the overt behaviour of our human subject—i.e. to the protocol—then it would constitute a good theory of the subject's problem-solving. (Newell and Simon 1963, 283, quoted in Shanker 1998, 70)

Shanker concludes that Newell and Simon mean that some correspondence

between mechanical processes (executable by computers) and human mental processes (realizable in biological brains) will facilitate viewing the problem-solving processes expressed via symbolic logic as a faithful representation of how human agents actually solve the problem, even if some of the steps of the human's problem-solving recipe are inaccessible. Shanker further points out that this is an idea originally put forward by "transcendental psychologists," who believe that the understanding of human mental activities should be in light of a psychological paradigm which is revealed transcendentally (Newell and Simon 1963, 293, quoted in Shanker 1998, 71).

Once the putative parallelism between symbolic AI and psychologism is in his sights, Shanker immediately introduces Wittgenstein's criticism of psychologism so as to undermine symbolic AI. The textual resource he relies on in this move is Wittgenstein's posthumously published work *Remarks on the Foundation of Mathematics* (1974, hereafter "*RFM*"), which, like *PG*, dates back to his complicated intellectual transition from *Tractatus* to *PI*. Quotations from *RFM* favored by Shanker include *RFM* I §116, wherein Wittgenstein asserts that the case in which the laws of inference compel an agent to think in a certain way or perform certain speech acts is not comparable to a case in which rails compel a locomotive to move on tracks (quoted in Shanker 1998, 98). He also cites *RFM* I §251, where Wittgenstein warns people not to explicitly formulate expressions like "$p \neq p$," since such expressions may misleadingly lead people to take "$p \neq p$" as an empirical statement—and if this were the case, both the affirmation and denial of it would be conceivable (quoted in Shanker 1998, 97). By placing these statements together, Shanker eventually delivers his "verdict" on the putative alliance between AI and psychologism:

> Psychologism is the consequence of construing such grammatical propositions experientially, thereby treating logic as 'a kind of ultra-physics, the description of the "logical structure" of the world, which we perceive through a kind of ultra-experience (with the understanding e.g.)' (*RFM* I §8). This last parenthetical remark makes it clear that Wittgenstein intended his investigations into the nature of logic to be seen as applying to the epistemological tradition extending from Kant all the way up to Russell (whose ill-fated theory of 'logical experience' in *Theory of Knowledge* was one of Wittgenstein's primary targets in the *Tractatus*): an epistemological tradition which applies as much to the cognitive significance attributed to formal models by AI as it does to the nineteenth-century debate over the foundation of logic. (Shanker 1998, 102)

How to re-evaluate this verdict? As I mentioned above, there are two issues involved here: one is whether Wittgenstein is actually criticizing the so-called

"epistemological tradition extending from Kant all the way up to Russell," i.e., a tradition labeled as "psychologism" by Shanker; the other is whether this tradition is inherited by modern AI. I have no further comments concerning the first topic, since a battle between Wittgenstein and psychologism is not so relevant to AI if the correct answer to the question implied by the second topic is negative. I argue that this answer should indeed be negative, and have two reasons for doing so.

First, as the testimony offered by Newell and Simon demonstrates, symbolic AI routinely employs programs compiled in accordance with symbolic logic to simulate the mental processes of human agents, thereby making AI itself a discipline bearing an affinity with cognitive psychology. This does not, however, mean that the validity of logic is thereby confined to the sphere of psychology. A more comprehensive characterization is that logical rules merely constitute some of the technical premises of symbolic AI, whereas its other premises are induced by empirical conjectures on how a mind is expected to function properly. That is to say, the resulting program should be viewed as a synthesis of its logical component and psychological component, rather than as a conversion of logic into psychology, which may eventually result in psychologism. Shanker's confusion of the two cases, in my view, is based on his misunderstanding of the relationship between theoretical subjects like logic and applied subjects like AI. The employment of the former in the sphere of the latter does not imply any reduction of the former to the latter. Otherwise the application of, say, mechanics to ballistics would also imply a retraction of the independence of mechanics itself, which would be patently wrong. In this sense, scholars working either in the field of ballistics or AI do not need to justify their respective theoretical premises, since this work should be handled by pure theorists working on higher levels if a framework of "division of labor" is provided in advance. Hence, by keeping the very idea of a "division of labor" in mind, one can easily see the sense in which Newell and Simon's work in AI diverges from classical psychologists' attempts to root logic in psychology, and thereby understand why the criticism of psychologism raised by Husserl, Frege, and (perhaps) Wittgenstein is not connected to our current concern.

Second, Newell and Simon's statement dates to a time when both AI and cognitive psychology were first emerging, and when the two disciplines were widely believed to be mutually supportive. Only in such an intellectual atmosphere could it seem natural to view the enterprise of AI as a simulation of human mental processes. The prevailing intellectual atmosphere has changed significantly since then. Indeed, most AI scientists today, who may not specialize in cognitive psychology, are more inclined to view cognitive psychology as *one of the* potential "wisdom banks" from which they might draw inspiration; for them, cognitive psychology has no special status, and must compete with other

fonts of intellectual inspiration which may be entirely irrelevant to psychology. Meanwhile, the goal of achieving a global simulation of human minds has faded away in most AI researchers' agendas, replaced by that of building an intelligent machine however possible, regardless of whether or not it simulates human cognition. Indeed, many representative AI approaches today do not simulate human minds at all. Take Bayesian networks (systematically introduced in Pearl 1988; 2009) as an example: they are probabilistic graphical models that represent a set of variables (each representing a type of event) and their conditional dependencies (each representing a causal relation between two events), and varieties of probabilistic inferences are expected to be facilitated in such models. It is noteworthy that the theoretical foundation of this approach lies in probability theory and graph theory, neither of which corresponds to any psychological reality of the human mind. On the contrary, there is an abundant supply of psychological evidence showing that human subjects are inclined to commit certain typical logical or probabilistic fallacies (cf. Piattelli-Palmarin 1994), and most AI scientists believe that such fallacies should be strictly avoided by AI. Hence, Shanker's picture of the putative alliance between AI and psychologism seriously neglects the recently emerging gap between the two. He is attacking a straw man.

5 Concluding Remarks

I have hitherto scrutinized Shanker's three arguments for identifying Wittgenstein as a foe of AI: the argument from Wittgenstein's own statements, the argument from the "rule-following" feature of language-games, and the argument from the putative alliance between AI and psychologism. As I have shown, none of these arguments work. However, due to the limitation of space, not all of the arguments in Shanker (1998) have been covered in this article. Shanker's remaining arguments include that AI cannot do justice to either the role that metaphors and insights play in scientific discoveries (cf. Ch. Four), or to the vagueness, ambiguities, and flexibility embedded in human's conceptual systems (cf. Ch. Five), whereas these factors are all substantially integrated into the philosophical ideas of Later Wittgenstein, e.g., in his famous account of the nature of concepts in terms of "family resemblance." These arguments, in my view, are, however, even less convincing than the previous three insofar as they are all established on an apparently false premise: that from the observation of what computers cannot do now, we can conclude what computers can never do.

My conclusion that Wittgenstein is not a foe of AI does not however entail that Wittgenstein is a friend of AI, a characterization which, if true, would be more

directly supportive of a desired interdisciplinary cooperation between philosophy and AI. Obviously, to go further and make Wittgenstein into a friend of AI requires further work and lies beyond the scope of this article, but such an attempt is worth making. Developments in the field of cognitive linguistics have already led us to see how Wittgenstein's philosophical ideas concerning "family resemblance" nourish the so-called "prototype theory of concepts" (cf. Medin & Wattenmaker 1987), and AI scientists can surely follow their lead.

Some may still contend that Shanker's picture of the Wittgensteinian position on AI may be metaphilosophically motivated by a broader view concerning the relationship between Wittgenstein and natural science, or, to put it more explicitly, Wittgensteinians taking this view need to work on a conceptual level which is above all empirical inquiry, as any attempt to blur the distinction between the two levels will inevitably lead to a surrender to scientific naturalism and hence the end of philosophy's independence. Insofar as I know, this anti-naturalistic mood has also infected Hacker (1996), who depicts post-Wittgenstein analytical philosophy (which is more scientifically oriented) as a tragic deviation from the Wittgensteinian metaphilosophical route. (In this sense, Shanker's work can be simply viewed as a more directed refinement or development of Hacker's view.)

A strict adherence to this metaphilosophical picture is simply a mark of stubbornness. In brief, I have two complaints about the so-called "anti-naturalistic trend" among Wittgenstein scholars. First, even if we grant that Wittgenstein did stress the science-philosophy distinction, this implies neither that a scientific mind cannot be inspired by his philosophical ideas nor that Wittgenstein himself is expressing hostility to empirical inquires of certain type. (To say that *A* is distinct from *B* does not mean that an *A*-*B* connection is in principle illusive.) Hence, there is still a large space in which a more science-oriented reconstruction of Wittgenstein's thought can be built, if this re-working itself is intended to be more beneficial to science rather than to historical studies of individual philosophers. Second, the current anti-naturalistic style of Wittgenstein studies can hardly make them fruitful and constructive rather than merely critical and negative, since the desired fruitfulness and constructiveness are usually brought about as a by-product of interdisciplinary inquires which bring philosophical ideas down to earth. The present prosperity of the philosophy of psychology provides a perfect gloss for saying so, so why can't the philosophy of AI follow this pattern by providing more glosses of this type? Hopefully Wittgenstein's wisdom can be more positively used by more open-minded Wittgenstein scholars to make a more fruitful relationship between philosophy and AI possible rather than to undercut it (cf. Xu 2014 as a tentative proposal for doing this).

Acknowledgments This research is sponsored by the National Social Science Fund of P.R. China (Grant No. 15ZDB020).

References

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York and Oxford: Oxford University Press.

Descartes, Rene. 1637. *Discours de la méthode*, in *Philosophical Essays and Correspondence*, edited by Roger Ariew, 46–73. Indianapolis: Hackett Publishing Company.

Dreyfus, Hubert L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason* (Revised Version). Cambridge, MA: The MIT Press.

Garson, James. 2010. "Connectionism," *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), edited by Edward N. Zalta, accessed 18 April 2015, <http://plato.stanford.edu/archives/win2012/entries/connectionism/>.

Hacker, P. M. S. 1996. *Wittgenstein's Place in Twentieth-Century Analytical Philosophy*. Oxford: Blackwell Publishers Inc.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: The MIT Press.

Hobbes, Thomas. 1651. *Leviathan, or the Matter, Form & Power of a Common-Wealth Ecclesiastical and Civil*, edited with an introduction and notes by J. C. A. Gaskin (1996). New York: The Oxford University Press.

Kiverstein, Julian & Michael Wheeler, eds. 2012. *Heidegger and Cognitive Science*. London: Palgrave Macmillan.

Knuth, Donald. 1976. "Mathematics and Computer Science: Coping with Finiteness." *Science* 194.4271: 1235–42.

Leibniz, Gottfried Wilhelm. 1714. *The Monadology*, translated by Robert Latta in 1999, accessed 18 April 2015, <http://www.rbjones.com/rbjpub/philos/classics/leibniz/monad.htm>.

Medin, D.L. & W.D. Wattenmaker. 1987. "Family Resemblance, Conceptual Cohesiveness, and Category Construction." *Cognitive Psychology* 19: 242–79.

Neumaier, Otto. 1987. "A Wittgensteinian View of Artificial Intelligence," in *Artificial Intelligence*, edited by R. Born, 132–73. London: St Martin's Press.

Newell, Allen & Herbert Simon. 1976. "Computer Science as Empirical Inquiry: Symbols and Search." *Communications of the Association for Computing Machinery* 19: 113–26.

Newell, Allen & Herbert Simon. 1963. "GPS: A Program that Simulates Human Thought," in *Computers and Thought*, edited by Feigenbaum, E.A., and Feldman, J. New York: McGraw-Hill.

Palmarini, M. Piattelli. 1994. *Inevitable Illusions: How Mistakes of Reason Rule Our Minds*. New York: John Wiley.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann Publishers.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference* (Second Edition). New York: Cambridge University Press.

Penrose, Roger. 1989. *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics*. New York: Oxford University Press.

Penrose, Roger. 1994. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. New York: Oxford University Press.

Searle, John. 1980. "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3.3: 417–45.

Searle, John. 2004. *Mind: A Brief Introduction*. New York: Oxford University Press.

Seidel, Asher. 1991. "Plato, Wittgenstein and Artificial Intelligence." *Metaphilosophy* 22.4: 292–306.

Shanker, Stuart. 1998. *Wittgenstein's Remarks on the Foundations of AI*. Abingdon: Routledge.

Wittgenstein, Ludwig. 1921. *Tractatus Logico-Philosophicus*, translated by D.F. Pears and B.F. McGuinness (1961). London and Henley: Routledge & Kegan Paul.

Wittgenstein, Ludwig. 1934. *Philosophical Grammar*, edited by Rush Rhees, translated by Anthony Kenny (1974). Oxford: Basil Blackwell.

Wittgenstein, Ludwig. 1935/1958. *The Blue and Brown Books*. Oxford: Blackwell Publishers Ltd.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*, 4th edition, translated by G.E.M. Anscombe, P.M.S. Hacker and Joachim Schulte (2009). Oxford: Wiley-Blackwell.

Wittgenstein, Ludwig. 1983. *Remarks on the Foundations of Mathematics*, revised edition, edited by G.H. von Wright, R. Rhees and G.E.M. Anscombe, translated by G. E. Anscombe. Oxford: Basil Blackwell.

Xu, Yingjin. 2013. *Xinling, Yuyan he Jiqi: Weitegensitan he Rengong Zhineng Kexue de Duihua* 心灵，语言和机器：维特根斯坦哲学和人工智能科学的对话 (*Mind, Language and Machine: The Dialogue between Wittgenstein and Artificial Intelligence*). Beijing: Renmin Chubanshe.