I would like to publish a series of five articles in February 2025 concerning the epistemological implications of reverse-brain engineering. In these articles, it will be shown that, contrary to what most philosophers allege, epistemology can and should be “naturalized”, i.e. done through empirical examination of actual cases of knowledge-acquisition as opposed to through a priori conceptual analysis. Importantly, this is subject to the qualification that the systems being empirically examined would be AI, as opposed to human beings or other living creatures. We can only speculate as to the nature of human cognitive processes; but we know exactly how AI functions---exactly what algorithms and protocols it follows---since we ourselves create it. Moreover, whereas our hypotheses concerning human cognition tend to be extremely vague and therefore hard to evaluate, the protocols governing AI are maximally clear---as they must be, since machines cannot respond to vague mandates. Now, there exist problem-sets that are not only large but unbounded (i.e. capable of expansion *ad infinitum*) such that both human beings and AI can solve the members of those problem-sets. Given such a problem-set, the protocols that AI uses to solve its members must coincide with those that human beings use to solve them, the reason being that any two protocols that generate the solutions to an unbounded problem-set necessarily collapse into each other. The problems discussed in this paper-series include: inductive-inference, data-modeling, language-learning, the nature of unconscious activity, musical creation, and rationality in the context of making predictions about the stock-market.

Article 1 will discuss what is involved in naturalizing epistemology. It will be argued that reverse-engineering (or “replication”, as I sometimes refer to it) is an epistemic modality in its own right---involving but not identical with data-modelling--and that we cannot possibly understand the mind except through this modality. Stated as clearly as possible: We cannot, even in principle, understand human cognition except through reverse-engineering.

In Article 2, philosophical views concerning the nature of non-analytic inference (e.g. inference from the past to the future, from the macroscopic to the microscopic, and from the observed macroscopic to the non-observed macroscopic) are scrutinized. It is shown that by constructing AI that can perform such inferences, we can understand both how we make them and also why they are sometimes knowledge-conducive.

In Article 3, it will be shown that, given AI that has already been constructed or is in the process of being constructed, Chomsky’s theories concerning language-learning and -comprehension can be evaluated. It will be seen that, relative to what existing AI indicates, those theories are likely correct.

In Article 4, it will be shown that we can understand the cognitive processes involved in music-composition on the basis of music-composition AI. In the process, the relationship of musical to other forms of cognition will be clarified.

In Article 5, we will discuss the American stock-market. We will show that, although it is inherently unpredictable in many respects (this being why day-traders tend to lose money), there are a few respects in which it is predictable (this being why long-term investors tend to make money); and we will explain how to construct AI that uses this information to maximize the returns on American stock-options (stock-options that can be exercised before expiration).

We will show that epistemology can be ‘naturalized’, meaning that it can be profitably studied by studying actual cases of knowledge-acquisition, as opposed to merely engaging in armchair conceptual analysis. According to the philosophical conventional wisdom, epistemology *cannot* be naturalized. The reason given is that empirical research can tell us *what* is involved in knowledge-acquisition, but it cannot tell us *why* what is involved leads to knowledge, it being the very purpose of epistemology to give us this information.

This reasoning is spurious on several levels. First of all, there is a difference between describing how people generate beliefs in general and how they generate knowledge-constitutive beliefs. Obviously, we cannot read epistemic norms off of the processes involved in belief-formation, since they often violate those norms, or only comply with them accidentally, and therefore fail to lead to knowledge. But the story is very different if we focus only on cases of belief-formation where the belief in question is indeed a case of knowledge; for in such a case, the relevant norms are indeed complied with, and for the right reasons, and they are therefore coded into those processes. Of course, other people’s thought-processes cannot be known directly; indeed, one cannot directly know one’s own thought-processes except on the rare occasions that they are conscious. Nonetheless, *if* we had direct knowledge of a knowledge-acquisitive thought-process, we would find it to embody due sensitivity to the relevant epistemic norms. When philosophers say that epistemology cannot be naturalized, what they are really saying, though they may not know it, is that epistemology cannot be naturalized *given* that we can only know of the operative processes through intermediaries that, unlike those processes *per se*, do not encode the relevant epistemic norms. If I watch someone solve a math problem, I see bodily movements, not thoughts, and I therefore don’t see how those thoughts are hewed to the relevant epistemic norms.

So, given that we cannot possibly ever acquire direct awareness of other people’s thought-processes, how can epistemology be naturalized? The answer lies in the fact that we can *build* systems that are functionally intelligent, if not actually so. Given an intelligent system that we ourselves have constructed, we know exactly what it is doing when it solves a given problem, since we were the ones who told it how to do so. The obvious counterobjection is that, given only that an intelligent system solves a given problem in a given way, it doesn’t follow that *we* would solve it in that way. And that is true with respect to isolated cases of knowledge-acquisition: there are often multiple ways of solving a given problem, and it therefore cannot be inferred that two distinct entities that solved it did so using the same methods. But when a system can solve multiplicities of problems, the range of viable hypotheses as to how it solved a given problem is much narrower, since any such hypothesis must be compatible with hypotheses as to how that system solved various other problems; and if that multiplicity is unbounded, then the range of such hypotheses is unlikely to be more than one. A comparison: If somebody picks a winning stock once, it could be luck, and if it isn’t luck, that person could have used various different methods. If somebody picks 1,000 winners in a row, then that person has a scalable strategy of some kind or other; and if that person picks winners in perpetuity, the range of such strategies is likely to be one. To close the argument: if we can build AI that solves an unbounded range of problems that human beings can solve, then it *is* likely that the AI using the same methods that we use, however unconsciously. And that likelihood is even greater if we can build AI that shares a person’s ability to solve each member of an unbounded *range* of unbounded problem-sets. Indeed, that likelihood is 100%, as we will later discuss. Consequently, epistemology can be naturalized by building AI that can solve unbounded sets of problem-sets that we too are able to solve.

There is another fact that is overlooked by philosophers discussing the possibility of naturalizing epistemology. Hypotheses as to how the mind solves problems tend to be extremely vague, and they also tend to be hard to falsify.

Xxx

According to Quine (1969), epistemology should be “naturalized”, meaning that, if we are to understand the nature of knowledge, we should study actual cases of knowledge, as opposed to engaging in *a priori* analysis of knowledge-related concepts. Philosophers either ignored or rejected Quine’s proposal, which consequently had no effect on the course of epistemology. In the present series of papers, it will be shown that epistemology can be naturalized, should be naturalized, and to some extent already *has* been. More specifically, it will be shown that the development of AI is largely nothing other than a successful naturalization of epistemology whose continued development will only deepen our understanding of knowledge---and not just in the trivial sense that we will know more about particular instances of knowledge, but also in the sense that we will acquire a deeper understanding of its very nature.

Reverse-engineering the brain: One can *directly* learn how mechanical objects work. I can learn how a given vacuum cleaner works by disassembling it, reassembling it, and otherwise directly engaging it. Nothing is comparable in connection with the brain. Indeed, the brain is the ultimate “black box”: we can know inputs and outputs, but we can only speculate as to the nature of the intervening machinery. Of course, we can study the brain; indeed, we can study both living and dead brains. We can study the brain’s anatomy, and we can also correlate changes in living brains with changes in mental activity. But such methods of study only go so far, for the simple reason that brains, unlike cars and vacuum cleaners, have both hardware and software.

Functionally identical systems that with identical configurations can run different softwares, and software-based activity is therefore inherently incapable of being understood strictly on the basis of physical configuration. We understand computers because we program them, and we don’t understand brains because we *don’t* program them. This does not mean that we cannot understand brains, but it means that we cannot use the usual methods of inquiry to acquire such understanding. Observation and experimentation are insufficient; a third operation must be added, namely, that of replication, otherwise known as reverse engineering. (We will use the terms “replication” and “reverse engineering” interchangeably.) Replication, we will see, is a distinctive form of empirical inquiry. It involves experimentation, observation, and hypothesis-formation, but it also involves much that isn’t involved in them. By replicating software-based systems, we learn how they *might* work; and when we triangulate this knowledge with knowledge of their configuration and behavior, we know how, in all likelihood, they *do* work.

Moreover, there is no other way to understand such systems. One does not first understand intelligence and on that basis create it; rather, one creates it and on that basis understand it. And, we will see, there is no other way to understand it, just as there is no way to understand color without having vision.

AI in connection with formal logic: There exists a discipline known as “formal logic” whose supposed purpose is to identify rules of inference---rules that say what can be inferred from what. Understandably, some logicians and philosophers went so far as to say that the purpose of logic was to identify the “laws of thought”, meaning, presumably, the laws followed by thought that is truth-conducive. So it was that the title of George Boole’s treatise on mathematical logic was one other than “The Laws of Thought”; and Leibniz, who attempted to mathematize logic, believed that he was thereby creating a “universal thinking machine.”

We’re all familiar with at least some of the laws of logic: if p, and if p entails q, then q (modus ponens); if p entails q, and if not q, then not p (modus tollens); if p entails q, and q entails r, then p entails r (transitivity): if everything has a given property, then an arbitrarily chosen object will have that property (universal instantiation); if each member of one class also belongs to some other class, and if each member of the second class also belongs to a third class, then each member of the first class belongs to the third; and so on.

There is a big problem with these so-called laws that many were quick to spot: following these rules doesn’t generate new knowledge. Of course, *violating* these rules involves making erroneous inferences; moreover, all legitimate inferences are *consistent* with them. But following these rules doesn’t lead to new knowledge. These laws certainly don’t tell us how to model data, i.e. they don’t tell us how to posit unknowns to account for knowns. But they do not even tell us how to reason in the vanishingly small minority of situations where the act of reasoning to be performed is to follow one of those very rules themselves. Suppose I know that is (a) x is taller than y and (b) y is taller than z. I cannot know that transitivity is applicable unless I already know that (c) x is taller than z. I know transitivity to applicable because I know that (c) follows from (a) and (b); I know that (c) follows from (a) and (b) because I know transitivity to be applicable. In general, one cannot know of *any* situation that a given law of logic applies to it unless one has already conceived of that situation that renders the corresponding inference unnecessary.

If a situation where it is simply stipulated that p is the case and entails q, minimal intelligence is needed to see that q must hold. But considerable intelligence is needed to see that a given situation not only holds but can be described by a proposition having a certain consequence. In a situation where it is simply a given that Jim owns three hotels and that Jim cannot possibly own three hotels without owning an odd number of hotels, minimal intelligence is needed to see that, indeed, Jim owns a prime number of hotels. But considerable intelligence is needed to understand those statements, and an even greater quantity of intelligence, albeit of a different kind, is needed model perceptual data in such a way that one knows that Jim owns three hotels. Formal logic of its very nature is idle until all of these bridges have been crossed, and it is useless after they’ve been crossed, for much the reason that an arithmetic lecture would be useless to someone who had to build design and use an airplane to attend that lecture.

By the beginning of the 20th century, philosophers were quite aware that the so-called laws of logic, while being worthy *objects* of reasoning, were of little use to reasoning itself. At this point, they did an about-face and claimed that the laws of logic not only weren’t laws of thought but weren’t even embodied in any significant way in thought; and they coined a term, “psychologism”, to denote the alleged fallacy of believing otherwise. What these philosophers did not say is that when they were referring to “logic”, they were referring only to certain specific logics---to Boolean algebra and the like. They did not consider the possibility that the specific systems of logic that had been developed were useless in the way of guiding thought, this might not be true of all such systems. having given up on the idea that logic of some kind or other might describe or guide thought, philosophers and mathematicians decided that logic’s rightful purpose was to delineate the foundations of mathematics, with some of them going so far as to claim mathematics to be *identical* with logic, either explicitly or implicitly. This program also failed, granting that it generated some useful results along the way. We will later discuss the reasons for this failure, since they are relevant to the present inquiry.

*Continued*: xxx