Artificial Intelligence, Machine Learning, and Human Beings

In a conversation with HackerNoon CEO, David Smooke, he identified artificial intelligence as an area of technology in which he anticipates vast growth. He pointed out, somewhat cheekily, that it seems like AI could be further along in figuring out how to alleviate some of our most basic electronic tasks—coordinating and scheduling meetings, for instance. This got me reflecting on the state of artificial intelligence. And mostly why my targeted ads suck so much...

AI is only as good as it's training set. The best AI is likely to be the AI which has had the most data points to compare and compute. Skinner and Pavlov's behaviorism made a similar assumption about the nature of human learning.

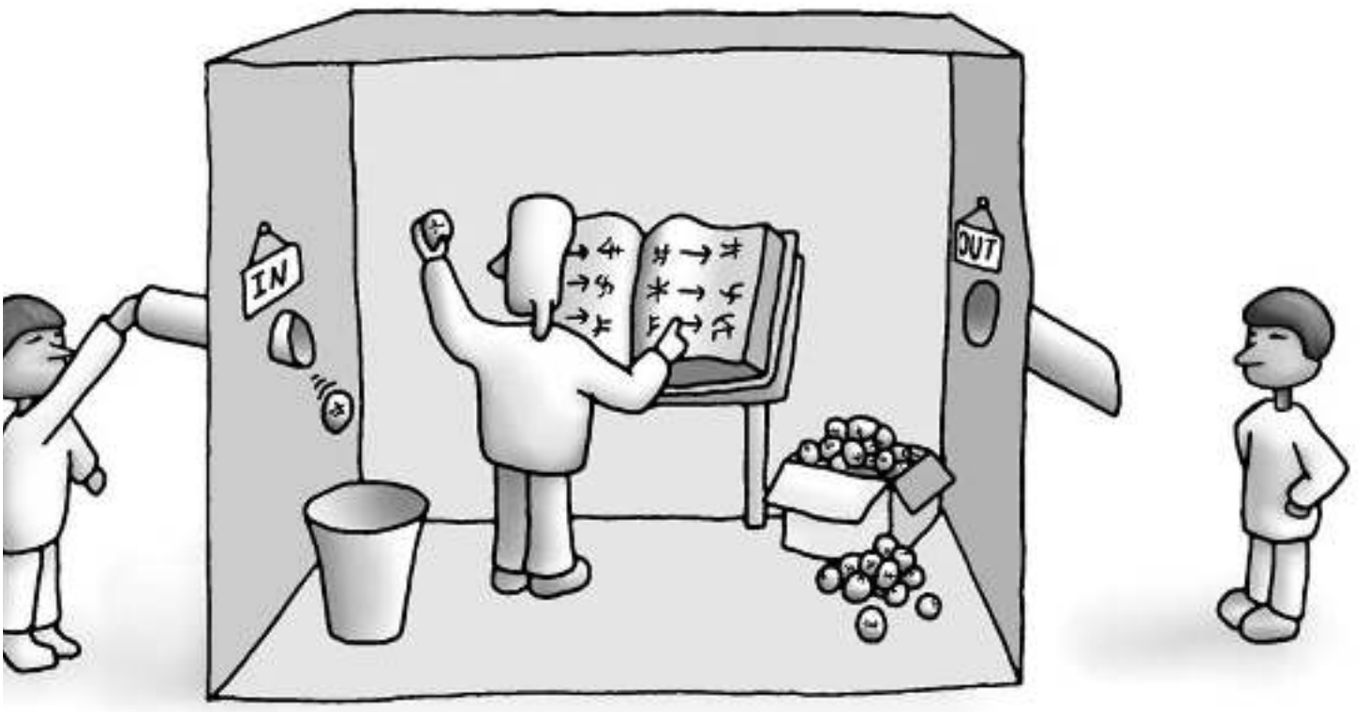
Namely, that humans are trained based on datasets, and the innerworkings of the mind are somewhat irrelevant. The human could simply be administered rewards and punishments as a condition of their output (behavior) and by this way would eventually achieve some desired target behavior.

The trouble is, that while psychology has largely moved on from this conception of human learning, machine learning cannot. This may represent a growing problem in the human-computer analogue and may give us a sense as to the recurrent problems AI architects can expect as they try to imitate human intelligence.

Importantly, humans seem to interpret and make meaning of their environments. In contrast, the computer is just a matching machine. Philosopher John Searle pointed out this difference long ago, by way of his Chinese Room thought experiment. In this, he illustrated himself in a closed room.

He does not speak nor understand Chinese. But, in this room he has a vast volume which contains a series of if-then statements. Through a slot on the door, he can receive strings of Chinese characters, look up the strings in the manual, and find the string of characters to send back out through the door slot. Clearly, Searle may convince people outside of the room that he really does

understand Chinese. But he does not; he is only matching symbols. And so his ability to appear knowledgeable is limited by the size of his Chinese if-then statement manual.



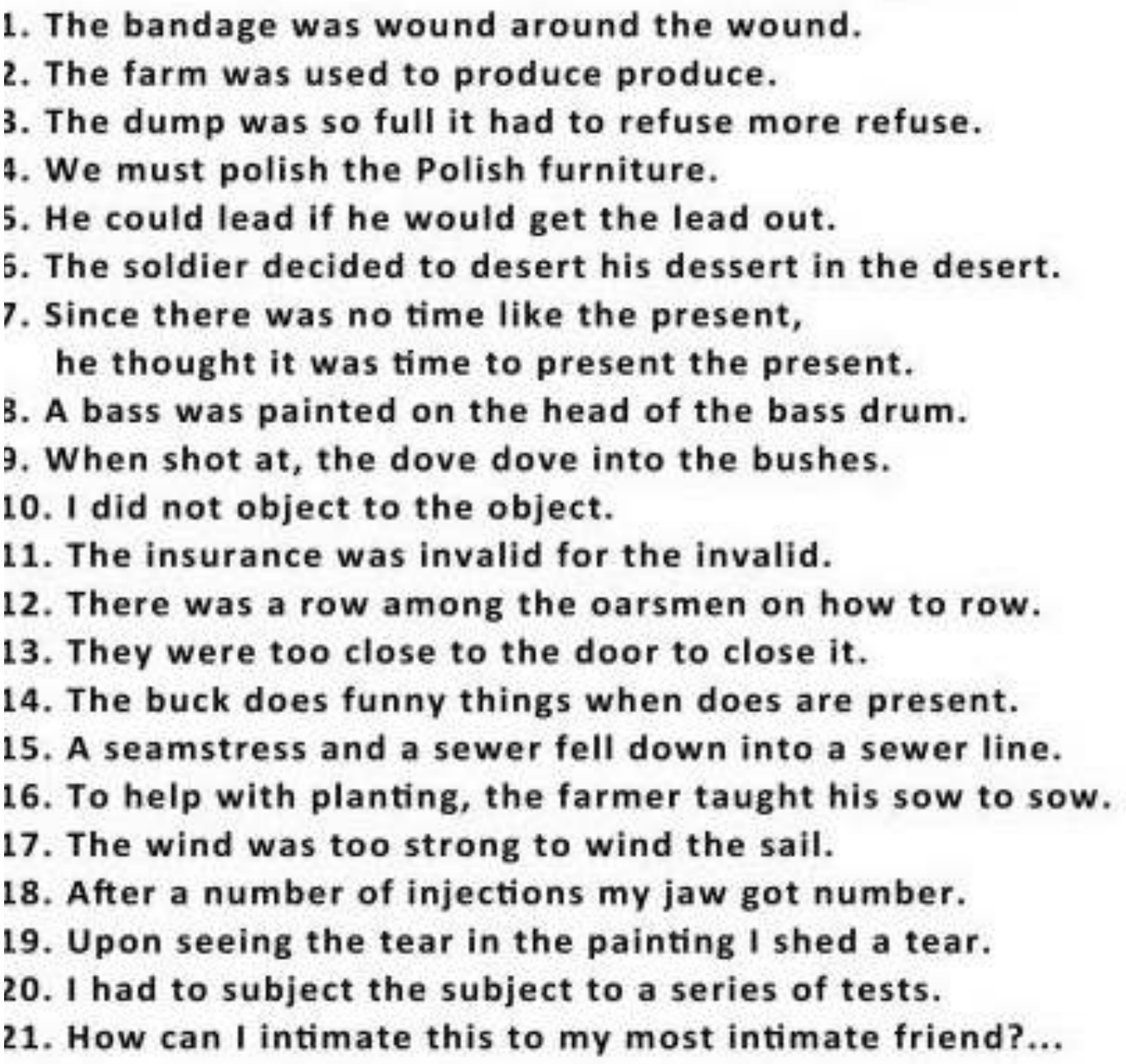
Some AI creators might claim that our meaning-making really is just an illusory byproduct of our large memory and quick matching capacities. This is the same type of claim that behaviorists made —that the mind is a black box, and learning can happen automatically, without some mystical sense of a ghost in a machine.

My intention here is not to claim that one must approach the mind mystically (I'd be out of a job as a psychologist if I did), but rather to point out where this reductionism went astray for behaviorists, and where it is therefore likely to become a pain-point for AI development.

Before discussing some of the experiments that eventually undermined machine learning-like models of human learning, I'd like to back up to address some very basic concepts. Computer "intelligence" is built from an array of binary switches.

And so, all things are self-contained entities which have no bearing on other things. For example, "good" means precisely "good." It carries with it no context; it exists in a vacuum of meaninglessness (much like the Chinese characters are to Searle). Thus, relationships between objects are not automatically captured by machine learning processes.

This is part of the challenge with natural language processing, in which the context of words co-constitute the meanings of one another (see graphic below for such examples). This ability is not natural to the computer. Thus, a programmer must manually write code for the machines to go out and "automatically" learn such relationships.



Human reason is phenomenologically different. Human reason is not binary. And so, things are not contextless self-contained units, but rather things reach out beyond themselves and exist in a network of relationships with other things. For instance, the meaning of “good” automatically implies “not bad.”

For the human, this oppositionality affords them the innate ability to distinguish between strings of different types, such as “good bad” and “good shoe.” For the computer, however, these oppositional relationships must be purposefully learned, and so “bad” is no more conceptually related to “good” than “shoe” is.

This may not seem all that important at first glance, but imagine if a human likewise thinks good:bad and good:shoe are equivalent statements. We’d assume they lack comprehension of either “good” or “shoe.”

In his book Artificial Intelligence and Human Reason, Joseph Rychlak, a psychologist for whom I have great respect, discusses such differences between computer and human reasoning. In it, he reviews studies which undermined behaviorist (AI-type) models of human learning. I will summarize this review here, and those interested can find the full review in his chapter “Learning as a Predicational Process."

In 1955, psychologist Joel Greenspoon tested the power of contingent reinforcement to operantly condition people toward certain behaviors. Specifically, Greenspoon asked participants in his study to say any words that came to mind out loud, one at a time.

For a 10-minute period participants listed words without any verbal reinforcement, which served as an experimental control. Following this period, the participants continued to list words and Greenspoon offered verbal reinforcement (an “Mmm-hmm”) each time the participant said a verbal noun. 10 of his 75 participants caught on to what the study was about and were excluded from the analysis.

In the remaining 65, Greenspoon claimed to have found automatic, unconscious behaviors conforming to the reinforcement (i.e. people started to list more verbal nouns, showcasing unconscious “learning”).

This type of a study should be very exciting to the AI developer who hopes that the principles of machine learning are essentially the same principles of human learning. No internal mental world was supposedly necessary to learn; the participants seem to have learned via pure association. If this is true, then the central problem of machine learning is how much training data can be thrown at it. But follow-up studies paint a far more complex picture.

In 1961, researcher Dulany revisited the knowledge levels of the participants in such experiments. He found that while many participants could not correctly state that the study was about learning to list plural nouns, many participants had developed “correlated hypotheses” which led to voicing “correct” words, such as “giraffes,” without technically having the right algorithm. For instance, the participant may have developed the hypothesis that they are supposed to list animals, and thus they started to say “giraffes, hippos, parrots, lions.”

These sorts of words would lead the experimenter to say “Mmm-hmm,” and to mark off that the participant had unconsciously learned (they could not correctly identify what the study was about yet were saying plural nouns). But these correlated hypotheses clearly are an indication that participants’ conscious hypotheses directed their responses, and that their thought process lead to the supposed “learning.”

Other follow-up studies, conducted by Page in 1969 and 1972 emphasized that cooperation of participants is another important factor in such studies. Page found that some participants actually were oppositional in their behaviors.

In his own replication of Greenspoon's work (in which he said "good" rather than "Mmm-hmm"), he found that some participants fell below chance (below their own base rate) in saying plural nouns in the second phase of the study (the phase in which reinforcement is applied). For example, participants who listed verbal nouns 20% of the time in the first phase without reinforcement to do so, might start to list verbal nouns 2% of the time after reinforcements were introduced.

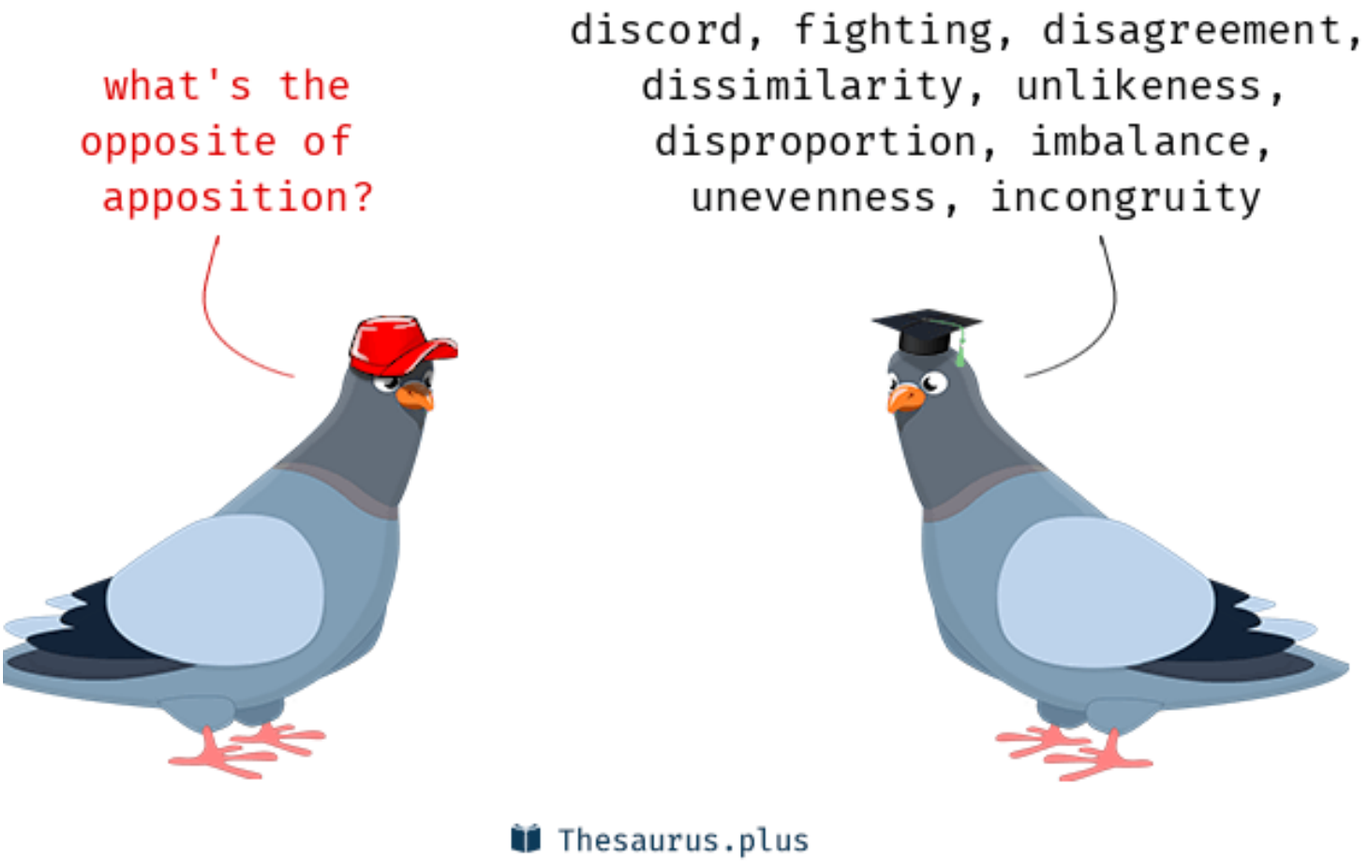
As the study developed, Page began to address these instances more directly. Once the participant's awareness of the rules and their level of cooperation were established, Page started specifically asking uncooperative participants to "make me say 'good'" and cooperative participants to "make me stop saying 'good.'" These participants immediately changed gears, and were able to make Page start saying good or stop saying good.

There were some uncooperative stragglers who continued to be uncooperative, and in such cases the participants were trying to abstain from what they believed to be unethical behavior on the part of the researcher (they thought the researcher was trying to influence the data to come out in some hypothesized manner, and didn't want to take part in helping manipulate the results).

Clearly, the results of these and similar follow-up studies are troubling for AI enthusiasts. For humans, it seems that some sort of a predicational process is taking place—people are making hypotheses about their world in order to understand it and to guide behavior.

Moreover, these hypotheses exist in relationship to alternative possibilities which allow for a mental flexibility that machines are not yet capable of. Page needed only to say "make me stop saying good" in order to completely reverse participant behavior. For AI, such a simple statement would involve a new training session. Past data cannot be instantly re-interpreted to derive the algorithmic corollary. This is in part due to the nature of oppositionality in human interpretation mentioned previously.

A machine does not intrinsically understand that "make me say good" and "make me stop saying good" share a special, oppositional relationship. A human cannot help but understand these statements as such. This is so deeply the case, that we even saw outright defiance on the part of some uncooperative participants, when their oppositionality equipped them to interpret a greater ethical authority than the authority of the experimenter. Such an intelligence makes one think of iRobot's Sonny, the robot who was far more "intelligent" than the others because of his human-like ability for understanding opposition.



Even further, there is evidence that human memory is quite different than machine memory. A machine, barring hardware malfunctions, memorizes contextless pieces of information. For humans, however, memory is dependent on the participant's ability to place it in a meaningful context. Craik and Tulving, for instance, found that when people are asked "Is _____ a type of fish?" they more easily remember the fill-in "a shark" than they remember the fill-in "heaven."

Likewise, people have an easier time remembering lists of things categorized by similarity than they remember lists without such categories. For computers, however, things are self-contained and are remembered without context. Making memories and retrieving them are irrelevant to conceptual or physical contexts, but this is not the case for humans.

As I've noted earlier, meaning-making is at the heart of the difference between humans and machines. This poses difficulties for AI in tasks related to identifying preferences, as well as for tasks involving judgment calls. A human naturally takes stock of their context to give meaning to their surroundings; considering a hammer to be contractor's tool in one context, a weapon in another, and a paperweight in yet another.

A person who is told "Seattle is not adjacent to Los Angeles on a map" may wonder, what if it was? They might fold their map is such a bizarre way as to bring Seattle and Los Angeles right next to each other, then settle into a sly, proud smile. Meaning-making arises from the combination of context cues and oppositionality; meaning-making is part and parcel with human reasoning. This is not the case for machines. At least not yet.

This is not to say that AI is doomed to failure. It's also not to ignore the incredible leaps and bounds that have already been made. I do not deny that machine learning is capable of mimicking any one human intelligence task very well. Any finite state machine is capable of this.

But to make artificial intelligence convincing, it's going to take a lot of edge-cases and alterations, re-training and re-training, and re-training. A ton of computing. AI as we currently know it is not flexible and cannot hope to be unless computers move from blindly matching objects to making

meaning. This leads me to the conclusion that AI, at least in the foreseeable future, is best suited to tasks which call for little conceptual flexibility, are relatively unaffected by context, and which are somewhat immune to human preferences (for instance, tasks such as aiding farming and flying aircraft). AI engineers should expect to nurse machine learning along for other tasks, such as making good recommendations (for food/movies/friends) or making ethical judgement calls (such as safety protocols for self-driving cars).

I'm helping build out the dWeb with ERA. If you enjoyed this article, you may want to check out my YouTube channel or connect with me on Twitter! <3