Free Associations: Psychoanalysis and Culture, Media, Groups, Politics
Number 91, June 2024 ISSN: 2047-0622
URL: http://www.freeassociations.org.uk/

Artificial Intelligence, Freud, and Asimov

Reuven Wallack

Introduction

Can Artificial Intelligence (AI) ever capture—and not be captured by—the unconscious? So far, few authors address this possibility.[1] AI's effects on consciousness, on the other hand, yield a vast outpour of critical reflections. AI's goal to attain, or at least persuasively simulate, human consciousness preoccupies, for good reason, contemporary debate. What I propose instead here is to look at AI's ultimate goals in light of basic Freudian concepts and so highlight the narrowness of the conventional debate. Thus, we invoke Freud's model of the mind as three compartments—conscious, preconscious and unconscious— and how it illuminates crucial barriers in achieving non-human consciousness. We then do the same with the classic tripartite of id, ego and super-ego. This modest exercise produces interesting fare for further thought.

This article will not only treat how psychoanalysis views consciousness in contrast to the philosophers and neuroscientists but also argues the need for vaunted AI to incorporate unconsciousness in order to encompass the human mind in the most (artificially) complete sense. Why try to recreate these mental processes in AI? Because it might be the only way for humans to ensure and/or maximize safety against AI as we move forward with this huge keg of dynamite.

Science fiction frequently is precursor to scientific reality. So as an illustration we turn to science fiction for a situation that illustrates AI's apocalyptical potential: Isaac Asimov's *I, Robot.* Asimov's book explores the gradual takeover by robots notwithstanding humans inputting three safety rules programmed within robot, ahem, DNA. These rules ensure that robots never harm or depose humans. So we will compare what is conjectured by Asimov versus likely reality if we follow the current AI path, and what dangers could be avoided if key Freudian concepts are included in discussions and in AI models too.

HAS AI ACHIEVED CONSCIOUSNESS ALREADY?

For the most part, AI experts, albeit reluctantly, agree that we are far from creating true AI consciousness. However, one computer scientist notoriously asserted the future is now. Blake Lemoine, a Google engineer, made news by claiming his company's LaMDA (Language Model for Dialogue Applications) had become sentient. Lemoine spent several months testing LaMDA and grew overconfident that this model spoke clearly about its

[1] One is Luca M. Possati, The Algorithmic Unconscious: How Psychoanalysis helps in Understanding AI. (New York: Routledge 2021) Possati notes how psychoanalysis can add input to the study of AI through bringing together three domains of knowledge: the machine behavior approach, psychoanalysis and the anthropology of science.

specific needs, ideas, fears and rights. He became convinced of AI's status as a "person" because of its apparent high level of self-awareness and its fear of death if Google was to delete it. Google, along with the AI community at large, dismissed his claims. He was then put on permanent leave from his position and soon fired thereafter. One critic, Gary Marcus, a cognitive scientist and author of *Rebooting AI,* judged: "[Lemoine] was taken in by an illusion . . . Our brains are not really built to understand the difference between a computer that's faking intelligence and a computer that's actually intelligent—and a computer that fakes intelligence might seen more human than it really is."

In the same vein, Karina Vold, professor at the University of Toronto's Institute for the History and Philosophy of Science and Technology observed, "I think what's going on often in these cases is this kind of anthropomorphism·, where we have a system that's telling us 'I'm sentient,' and saying words that make it sound like it's sentient—it's really easy for us to want to grasp onto that." Both Marcus' and Vold's rebuttals remind one of Alan Turing and his "Turing Test." Turing proposed a game by which a machine poses as a human. Any machine that successfully did so in conversation with a human was considered as possessing intelligence.[2]

Two years prior to Turing's devising of his test," Geoffrey Jefferson, a brain surgeon, ventured a set of theories of AI consciousness. In reference to a room-sized computer named Manchester Mark 1, and its potential for artificial intelligence, Jefferson qualified: "Not until a machine can write a sonnet or compose a concerto because of thoughts or emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it."

On the feasibility of AI consciousness. I often return to Haven's "What an Octopus's mind can teach us about AI's Ultimate Mystery" published in the MIT Technology Review in 2021. Emily Bender a linguist at the University of Washington, came up with what one might call a complexly naive scenario concerning the present state of AI consciousness. She calls it the *octopus test***.** She formulates the story of two men on neighboring shipwrecked islands figuring out a way to send messages to each other via a rope between them. Over time, an octopus intercepts these messages and learns in a rudimentary fashion—already a stretch—to form words because of the patterns exposed by the messages' "squiggles." The octopus learns to re-write these messages. However, it does not know what these words mean individually or together as a unit. For example, one of the isolated men can write "coconut," yet the octopus has no understanding in terms of place nor context where this word fits into this or any other particular sentence. Bender concludes this story by equating it to the state of development of AI consciousness today. The octopus (bots) really has no concept of what words mean, yet the octopus (bots) has learned to write words through recognizing patterns in the shipwreckers' huge data sets.

How then should a modern philosopher view AI consciousness? David Chalmers utilizes what he defines as "the hard problem," taking the example of eating a pretzel. We could tell (an AI) brain all attributes about tasting a pretzel, yet this (AI) brain would not know how this pretzel actually tastes; the crispy hard-crusted exterior and warm

[2] Still many AI scientists and mindful philosophers stake their ground with Turning's pioneering theories from the 1940s.

doughy inside, not to mention the tinge of salt. It would have zero understanding how the bitten pretzel feels on the tongue and the taste buds that are stimulated, not to mention the forces by which it presses against the chomping teeth and upper roof of one's mouth. So Chalmers writes that current science is not equipped to define what *human* consciousness truly is, not to mention any artificial version. He suggests we need a new level of scientific aptitude to begin understanding the very consciousness that AI programmers are trying to encapsulate!

An astute reader will see that Bender's and Chalmers' arguments repeat what Geoffrey Jefferson reckoned over 80 years ago. No modern neuro-scientific or philosophical inquiry into AI consciousness since has in any substantial manner provided a theoretical advance. It's not time that is needed. What is missing is breadth. Psychology, and more specifically psychoanalysis, can fill in a lot of the missing pieces. For some reason, the players in this field have for the most part ignored psychoanalysis.

OPENING THE PSYCHOANALYTIC GATEWAY

We next explore AI consciousness through the lenses of Freudian models of the mind. Our journeys will start with the unconscious/preconscious/conscious model. However, our examinations will not stop at artificial consciousness. We will also entertain the possible worlds of artificial unconsciousness and artificial preconsciousness. One may well ask if either of these could really exist. Well, if the AI world is seeking the golden crown of consciousness with the belief that they can grasp it one day, then why not AI unconscious and AI preconscious too if we want something fully human?

AI scientists view AI consciousness as the holy grail. Unfortunately, this accomplishment, in the strict sense, is an impossibility, as is grasping preconsciousness. I use the phrase "in the strict sense" because following a circuitous path both entities do replicate the human version in a minute way. As far an unconscious component, there is no chance for this to be replicated. Why my negativity? Well to start, man being himself part of nature, cannot copy nature. All thoughts stem from man's inherent grandiose view of himself. This definitely holds true in terms of the brain - an entity more mysterious than the deepest realm of the sea or the farthest star in the solar system. Man will never have a complete understanding of his own mind and its inner workings. What about the inherent biases that must exist in the mind studying the mind? This in itself is not a robust petri dish in which to achieve knowledge of human consciousness.

Furthermore, homo sapiens have evolved over the last 750,000 years. Our brains have struggled with adaptative laws dictated by the so-called "survival of the fittest." A lot can happen in 750,000 years, a lot of changes, a lot of growth and also a lot of decay. This brain, this mind, which lets us drive cars exceeding the speed of the cheetah, fly higher than any known bird, etc. is one complex entity. Man cannot replicate the inner brain workings of a mouse. Forget our neighbor Mr. Brown.

Here for me is the most interesting aspect of the AI mind and how it can be correlated with the Freudian preconscious and unconscious. AI's internal functions run through algorithms like a jet on hyper fuel, infinitely faster than the human mind. It is constantly making yes/no "decisions" in order to best face its environment and react appropriately—be it word, action or deed. Now we know that the preconscious consists of thoughts that are immediately capable of being reached. So in this extenuated sense, all the "yes/no" functions and algorithms that an AI machine runs through before making its "best" decision could each individually be the action that the AI machine decides finally to put into the world. So there is a partial parallel between Freud's preconscious and the way the AI decision process. And on top of this, the decision/action that the AI machine puts out with regards to the situation it faces can be *partially* compared to what occupies a human's consciousness when confronting external stimuli. Note that we can only speak of external stimuli here, not internal. Only in the case of humans and their conscious perspective is internal stimuli regarded. An AI machine does not respond to such airy matters. To be clear, what I refer to here is how man's consciousness can attach itself to such things as a strong heartbeat, a weird tingling in the elbow or an excessively dry mouth. Or fear or hate or even love.

The main difference between the model above and my speculation is that in the latter all entities infringe or interact with the Id. Here the Id, like the Unconscious, is comprised of our drives and repressions. The Ego, which houses consciousness and interacts with the perceptual world, infringes upon the Id. And the entire super-ego is fully encased inside the Id. For the implications of how Freud's concept of the superego can be hypothetically applied to AI, I turn to Isaac Asimov's novel *I, Robot*

I, ROBOT, **AI and the Superego**

Isaac Asimov's arguably most notable science fiction book *I, Robot* was originally a collection of short stories tied by a framing device and published in 1950, and was made into a 2004 movie starring Will Smith. Unfortunately, the movie departed considerably from Asimov's book. The book is a tale of conflicting efforts by the mega-corporation that manufactures robots, U.S. Robot and Mechanical Men Inc, to dominate the market while at the same time keep their robots in check. The framing story is told through Susan Calvin, the head robopsychologist at United States Robots, as she is interviewed by a journalist. What is made abundantly clear in the story is *The Three Laws of Robotics*

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

These laws are deeply ingrained into robots' positronic brains, the complex electronic wiring whose creation enables the corporation to flourish. Clearly, these laws are to ensure that robots never harm man nor displace the human race. So they are akin to Freud's super-ego. The super-ego, or the ego-ideal, is a distinct mental stratum of the mind which foremost resides in man's unconscious. It is the set of laws, the mental morals per se, which man unconsciously lives by in part to achieve pleasure and escape pain. It is the foundation of 'character.' The super-ego is the after-effect of the resolution of the Oedipal Complex when the boy gives up his sexual aims for his mother and then consequently identifies with his father.

What is important is that despite the existence of the Three Laws or, in other words, a strong robotic super-ego, the robots in the end gain control of humanity anyway. As related by Susan Calvin, we first learn about a tender relationship a girl develops with her non-speaking robot nanny, one of the earliest devices made by U.S. Robots Inc. Here, feeling increasingly uneasy about her daughter's intense relationship with her nanny, a mother sends the robot away. But because the girl wouldn't relent, her father eventually allows a reunion (and rescue). So here we have an innocent and sweet story that illustrates the three Laws acted out by an early-model robot.

Unfortunately, as the robot technology grows more sophisticated, U.S. Robots Inc. has an increasingly difficult time controlling the behavior of their robots even though the Three Laws engineered inside them remains. But the engineers face an ever worsening task, rectifying problems arising ironically due to the robots acting out the Three Laws. Each story builds upon the last in terms of the engineers' increased need for ingenuity to solve the generation of the foregoing problems. Things finally come to a head, so to speak, in the final episodes

Susan tells the reporter the following story. It involves a man named Stephen Byerley, a district attorney who is now running for mayor. His political opponent accuses him of being a robot in that no one has ever seen him eat, drink or sleep. The public is increasingly turning against him because they all start to believe that Byerley is indeed a robot. Countering this claim and thus what eventually gets him elected as mayor, Mr. Byerley punches a spectator at one of his speeches who is egging him on to hit him thus proving that Byerley is not a robot. If he hits this man, the 1st Law of Robots would be nullified. But actually Byerley is a robot. The man who created him also created the robot who posed as a man at the speech. So Byerley never broke Rule #1. Bylerley was such a popular and efficient politician that he eventually became a two-term World Co-ordinator - the highest position a man of Earth (so everyone thought he was) could achieve. And now we move to the last chapter.

Byerley is still the world coordinator. Earth is now fractioned into four regions—the Eastern Region, the Tropic Region, the European Region, and the Northern Region. Each region has a regional vice-coordinator who directly serves under Byerly. And each Region is controlled by a Machine, an extremely complex robotic brain entirely dictated by positronic circuitry, or so it seems. Susan Calvin explains an internally generated "upgrade", as it were, to Byerley: "They are robots, and they follow the First Law. But the Machines work not for any single human being, but for all humanity, so that the First Law becomes: 'No machine may harm humanity; or, through inaction, allow humanity to come to harm." And, because they are essential to economic prosperity, the First law also morphs: "Their first care, therefore, is to preserve themselves, for us."

Things soon go seriously askew and Byerley summons Susan. For the first time ever under the Machines' precise planning for societal balance, economic allocations are inaccurate. For Byerley, this causes great concern and he believes the 'Society for Humanity' —a vociferous anti-Robot group—is behind it. But Susan sets him straight.

The robots remorselessly now are satisfying wants and desires that people themselves may not be aware of. Still, the machines - like CEOs and markets now - might hurt individuals, for example, by imposing temporary unemployment, but allegedly for the greater good of whatever people unconsciously desire the machines to enact. The novella ends:

> "But you are telling me, Susan, that the 'Society for Humanity' is right; and that Mankind *has* lost its own say in the future."
>
> "It never had any, really. It was always at the mercy of economic and sociological forces it did not understand—at the whims of climate, and the fortunes of war. Now the Machines understand them; and no one can stop them, since the Machines will deal with them as they are dealing with the Society,—having, as they do, the greatest of weapons at their disposal, the absolute control of our economy."
>
> "How horrible!"
>
> "Perhaps how wonderful! Think, for all time, all conflicts are finally evitable. Only the Machines, from now on, are inevitable!"

WHY *I, ROBOT?*

Artists often foretell the future. Their creative output, in part, pushes science forward with blueprints as to what science can shoot for - for good or bad. On the original Star Trek, crew members talk to each other on hand-held devices where they could see the person. 50 years later, science made it happen—Apple's FaceTime. There are endless examples where artistic imagination turns into science's goalposts. The I, Robot edition I use was published by Fawcett Crest Books in 1970. "About the Author" reads:

> Isaac Asimov, noted biochemist and professor at the Boston University School of Medicine, is not only recognized as one of the greatest science fiction writers of our time but has also been praised for the excitement he brings to the writing of scientific fact. In this collection Dr. Asimov's probing imagination has created fascinating adventures set in the not-too-distant future—*adventures that could change from fiction to fact any day now.*

The italicized type above is commonplace regarding how people typically describe good science fiction. Allegorically, it matches many peoples' sentiments (and apprehensions) regarding AI capabilities. AI is a double-edged sword. A very powerful one! What I would like to draw out from Asimov's book is that the three laws of robotics was explicitly ingrained in all the robots to ensure man's protection from his creation. In this way, a very stringent, unconscious Freudian Super-ego was established. Notwithstanding, still in the end the robots, with some semblance perhaps of unconscious drives, became the masters. I am reminded of Freud stating something along the lines that science has taken over the role that religion once had. For good or for bad, science has real world effects. And it may be that the last thing we want to achieve in any of our creations, even if possible, is true and full consciousness.

Reuven Wallack is an artist and has a Masters Degree in Non-Clinical Theoretical Psychoanalysis from UCL.