AI AND FREUD'S MAPS OF THE MIND

INTRODUCTION

Throughout all my research on the subject of AI and its ability to capture the ephemeral unconscious, there are very few authors who explore this applicable entity.[1] On the flip side, with regards to consciousness much has been written. AI's goal to reach this highly sought achievement of consciousness which most say will push this technology to the next, or perhaps this same many or most argue to the penultimate level, it is the philosophers, neuroscientists, computer specialists and the like whose opinion that AI experts are trying to capture and capitulate. This does not make much sense to me. Is it not psychologists and psychoanalysts who most study and make their live's work to ascertain the healthy functioning of man's consciousness? With this in mind, what I propose here is to explore Freud's two concepts of the mind and see how his models apply to AI's ultimate goals of consciousness, unconsciousness and beyond. But why Freud's paradigms and not more modern psychologists or psychoanalysts? Because in truth, no better models of the mind have been established by his successors or in fact in any field of thought or science.[2] To even conjecture this we must remember that the idea of the mind is a very abstract one, so progress in its understanding travels a very unique path relative per se to the more even keel step-by-step increase of knowledge of particle physics. Thus, we are going to examine Freud's first model of the mind and how it is split it up into three compartments—the conscious, preconscious and unconscious and how it applies to the (perhaps ultimate) achievement of non-human consciousness and more. We then are going to do the same with Freud's later and perhaps more corrective model—the id, ego and super-ego paradigm.

This study will not only explore how psychoanalysis speaks about consciousness in contrast to the philosophers' and neuroscientists' take, but we will also go into the need of AI to capture man's unconsciousness among other Freudian mental processes to ultimately capture man's mind in the most artificially complete sense possible. But why try to achieve these mental processes in AI? Because in truth it might be the only way for us humans to ensure and/or maximize safety against AI as we move forward with this huge potential keg of dynamite.

We all know that science fiction frequently precursors scientific reality. So to end this paper, we will turn to science fiction and explore a situation, or similar there of, which illustrates AI's apocalyptical potential. We will turn to Isaac Asimov's *I Robot* to do this. Asimov's book explores the gradual take-over by robots over man notwithstanding man inputting three all-encompassing safety-measure rules programmed deep within the robots DNA. These rules

[1] One example is Luca M. Possati's extensive article "Algorithmic unconscious: why psychoanalysis helps in understanding AI." Possati speaks 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.

[2] I know this is a bold claim and of course unprovable. Yet, it is also not unprovable. We all know this claim is obviously strictly of conjecture. Yet, for this author never have I been presented a stronger model of the abstract mind.

should ensure that a robot's defeat over man could never occur. Yet, it does! So we will compare what is conjectured by Asimov versus what could be the reality if we follow the path current AI experts are on, and what dangers could be avoided if some Freudian concepts are included in present discussions and AI models themselves.

HAS AI ACHIEVED CONSCIOUSNESS ALREADY?

For the most part, AI experts agree that we are far from establishing AI consciousness. Though this achievement is no longer measured in years for now it is quite hard to gauge because the field is a brush fire of exponential scientific growth. However, one accomplished computer scientist thinks the future is now! Blake Lemoine, a Google engineer, made news recently by publicly claiming his company's LaMDA (short for Language Model for Dialogue Applications) had become sentient.[3] Sentient is another word for consciousness.

Lemoine had spent several months testing LaMDA and grew confident that this model spoke clearly about its specific needs, ideas, fears and rights. Lemoine became convinced of the AI's status as a "person" because of its high level of self-awareness and, for just one example of many, its fear of death if Google was to delete it. So much so that Lemoine came out to the public and proclaimed LaMDA had achieved consciousness. 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.

For example one of his detractors, Gary Marcus a cognitive scientist and the author of *Rebooting AI* said:

> "My view is that [Lemoine] was taken in by an illusion."

Marcus continues:

> "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 of Marcus' comments, Karina Vold, an assistant professor at the University of Toronto's Institute for the History and Philosophy of Science and Technology believes again Lemoine has been fooled:

[3] "A Google engineer says AI has become sentient. What does that actually mean?" Laura McQuillan, CBS News. 2022.

> "I think what's going on often in these cases is this kind of anthropomorphism[4], 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 statements very much remind us of Alan Turing[5] and his "Turing Test." The "Turing Test" is a piece of scientific philosophy from the 1940's that is regarded as the first major modern conjecture exploring if artificial intelligence is present or not. Quite simply, Turing proposed a game by which a machine tried to pose as a human. Any machine that successfully did so in conversation with a human was considered as having intelligence.[6]

HISTORIC VIEWPOINTS ON AI CONSCIOUSNESS

Two years prior to the "Turing Test" (which is relevant to artificial intelligence alone), another Englishman Geoffrey Jefferson, a brain surgeon, established the first groundbreaking theories indirectly referring to 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."

Here, Jefferson in his own way bespeaks not only of AI, but also provides the first taste of artificial consciousness.

[4] When one anthropomorphizes objects and creatures, one projects human-like characteristics that aren't really there.

[5] Alan Turing was a complete man of intellect and genius. Most notably he was an English cryptologist vital to the Allies ultimate success in World War II through deciphering German battle codes. Turing was also a great computer scientist and mathematician. He is widely considered the father of artificial intelligence through his development of the 'Turing Test.' Turing asked himself "Can Machines Think?" In all, Turing valued outside appearances and, more so, outside appearances of thinking above all else. One could argue that today's bots have achieved this measure already.

[6] Still today many AI scientists and mindful philosophers stake their ground with Turning's pioneering theories from the 1940's, even though the field of scientific mental cognition has grown by leaps and bounds in just the last few years.

Having now crossed in discussion the bridge between artificial intelligence (Turing per se) into artificial consciousness (Jefferson per se), we can now comfortably move forward and discuss more recent opinions of the feasibility and thorough potentiality of AI consciousness.

I have read many articles on AI conciseness, yet I continually return back to Haven's "What an octopus's mind can teach us about AI's ultimate mystery" published in the MIT Technology Review in 2021. I know that using an isolated research source for such an intricate subject is not the best of ideas. Yet with full knowledge in hand, I always return back to this one article. I believe Haven's article is fantastic in terms of subject matter and clarity. And I do believe if a reader is further interested in learning more about AI consciousness, this paper should be high on his list. It is not too long and it is clear and precise.

Let's now excavate from Haven's excellent paper:

To start with, Emily Bender a linguist at the University of Washington, has come up with a naively complex scenario concerning what she entails to be the present state of AI consciousness. She calls it the **octopus test.** She forms the story about how two men on neighboring shipwrecked islands figure out a way to send messages to each other via a rope back and forth between them. Over time, an octopus intercepts these messages and eventually learns in a rudimentary fashion to form words because of the patterns exposed by the messages' "squiggles." The octopus then learns to re-write these messages in such a way as to combine individual words together. However, it does not know in any way what these words mean individually or together as a unit. For example, one of the isolated men can write "coconut" in his message, yet again the octopus has no understanding in terms of place nor context where this word fits into this particular sentence nor any other. Bender concludes this story by equating it to the context of where present AI consciousness is at. 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.

What about how a modern philosopher views AI consciousness? David Chalmers from New York University, utilizes what he defines as "the hard problem." He takes the example of eating a pretzel. We could tell (an AI) brain everything and completely all attributes about tasting a pretzel, yet "this" (AI) brain would not know a thing in terms of what this pretzel actually tastes like. The crispy hard-crusted exterior and warm doughy inside, not to mention the exact tinge of salt. It would have zero understanding how the bitten pretzel feels on the tongue and the exact 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. In this way, Chalmers writes that current science is not even equipped to define what human consciousness truly is, not to mention the artificial version. He suggests we need a new complete level of aptitude and advancement to ever achieve this concept of consciousness to even begin incorporating what AI programmers are trying to encapsulate.

An astute reader will see that Bender's and Chalmers' arguments are really in truth what Geoffrey Jefferson put forth over 80 years ago. In general, no modern neuro-scientific, philosophical or other related field's inquiries into AI consciousness have in any substantial manner provided a theoretical advancement.[7]

However, there is one thing that Chalmers does do for us. It points out that modern science will have a hard time, or perhaps according to him, no chance of acquiring AI consciousness

[7] In fact, this is a major reason why I believe Haven's article can serve as a sole research outlet.

unless science significantly propels to a higher level of advancement so that it can even grasp a substantially improved understanding of the human mind. But 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 which I don't understand, the players in this field have for the most part ignored psychology and psychoanalysis and leaned on philosophical concepts of the mind and consciousness.

THE OPENING OF THE PSYCHOANALYTIC GATEWAY

And so we have finally arrived at the gateway of this paper. Our plan is to delve strictly into psychoanalytic theory. But why? Because past and modern psychology has been truly and mostly based upon action/reaction *modus operandi* therapeutics such as Cognitive Behavioral Therapy, Exposure/Response Therapy and Dialectical Behavior Therapy. Cheekily said, such is even the case with Maslow's dogs and the once popular Primal Scream! All these therapies and their associated theories offer no concepts of how the human mind, including consciousness, truly operates. And if they cannot do this, they are useless in examining this same entity in the artificial realm. Freud's psychoanalysis however answers the call.

As said in the beginning of this paper, we are going to explore AI consciousness through the lenses thrust upon us by Freud's two models of the mind. These being the unconscious/preconscious/conscious model and the superego/ego/id paradigm. Our journeys will start with the unconscious/preconscious/conscious model. However, our examinations will not stop at AI consciousness. We will also reach out to the possible worlds of artificial unconsciousness and artificial preconsciousness. One may ask themselves 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 onto it one day, then perhaps so does the entities of AI unconscious and AI preconscious exist.

APPLYING FREUD'S UNCONSCIOUS / PRE-CONSCIOUS / CONSCIOUS MODEL TO AI

As we pointed out earlier, today's AI scientists view AI consciousness as the holy grail. Unfortunately, this accomplishment in the strict sense is an impossibility. AI pre-consciousness holds this same sway. I use the term "in the strict sense" because following a circuitous path and one that is not truly intended 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 in any manner.

So why my negativity? Well to start off with, man being himself part of nature, cannot copy nature. All thoughts otherwise stem from man's inherent grandiose view of himself. This definitely holds true in terms of the brain…an entity more mysterious than both the deepest realm of the sea and the farthest star in the solar system. Man will never have a strong understanding of his own mind nor its inner workings. Plus, what about the inherent biases that must exist in the mind studying the mind.   This in itself is not a robust petri dish to achieve knowledge of the matter at hand.

Furthermore, generously speaking, homo sapiens have evolved over the last 750,000 years. From this date and quite definitely much before as applicable to homo sapiens' ancestors, our brains which are man's defining feature have struggled with adaptitative laws dictated by the "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 mode of machinery. Just for these reasons alone man will never replicate what Nature has made. Man cannot even replicate the inner brain workings of a mouse. Forget our neighbor Mr. Brown.

So in saying all of this and considering the immense shroud clothing man's unconscious, we immediately label AI unconsciousness as an impossibility.

What about Freud's pre-conscious? Here is the most interesting aspect of the AI mind and how it can be correlated with the Freudian mode in question. 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" to best face its present environment and react most appropriately—be it word, action or deed. Now, when we think of Freud's preconscious, we know that the preconscious consists of thoughts that are immediately capable of being conscious. So in this 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 ultimately decides to put forward into the world. In this sense, there is a partial parallel between Freud's preconscious and the way AI runs through its decision process. And on top of this, the action that the AI machine eventually puts forth with regards to the current situation that it then faces can in a stretch be **partially** compared to what occupies a human's consciousness when faced with a particular 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 matters.   To be clear, what I am referring to here are how man's consciousness can attach itself to such things as a stomping of a strong heart beat, a weird tingling in the elbow or an excessively dry mouth.   Or fear or hate or even love.

Though the analysis of this particular Freudian paradigm is quite short, it says all it has to say. We can now turn our attention to Freud's Id, Ego, and Super-ego model.

APPLYING FREUD'S ID / EGO / SUPER-EGO MODEL TO AI

The main difference between the model above and this one is that here all entities infringe or interact with the Id. Here the Id, like the Unconscious above, 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.[8]

Being that the Id/Ego/Superego paradigm is even more mapped through unconscious entities than the model examined above, we can only say that this paradigm has even less parallels with potential AI mind development.

However, before ending, there is one other item we would like to examine via literature to see how Freud's concept of the superego can be hypothetically applied to AI. We turn our attention to Isaac Asimov's great novel *I, Robot.*

I, ROBOT, AI AND FREUD'S SUPER-EGO

Isaac Asimov is one of the most highly regarded science fiction writers ever. His most notable book I, ROBOT was written in 1950. Of such acclaim, it was made into a 2004 movie starring Will Smith. Unfortunately, the movie did not hold strictly to the words of Asimov's masterpiece. As a book, I, ROBOT is a mind-spelling tale of a series of interactions between the quasi-monopoly who manufactures robots, U.S. Robot and Mechanical Men, Inc, and their attempts to keep their robots in check. The story is told through the eyes of Susan Calvin, the head robopsychologist at United States Robots, as she is being interviewed by a journalist.

What is made clear even before the Introduction of 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.**

Handbook of Robotics
56th Edition, 2058 A.D.

[8] For those not acquainted with Freudian theory, the superego houses the laws that a man learns from his interactions primarily with his parents and secondarily other earlier influencers such as teachers mostly in his early youth. Though most of it is phylogenetically determined by one's progress through the Oedipal Complex—the most significant psychic event in an individual's life.

The robots have these laws deeply ingrained into their positronic brains, this being the complex electronic wiring whose creation leads U.S. Robots to be the biggest company of their times. Clearly, these three laws are to ensure that robots will never harm man and, but more importantly never take over the human race. So in a sense they are very akin to Freud's concept of the super-ego which Freud writes about extensively in his article "The Ego and the Id." The super-ego, or the ego-ideal as it is also called, is a distinct mental strata 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, as Freud says, 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. Freud writes:

> "This leads us back the origin of the ego ideal; for behind it there lies hidden an individual's first and most important identification, his identification with the father in his own personal prehistory."

Returning to the robots, what is important about all of this is that no matter the existence of the 3 Laws, or in other words a strong robotic super-ego, the robots in the end end-up controlling the human race. I will speak more about the ramifications of the conclusion of this brilliant novella later, but first a quick summary of Asimov's book. Again as spoken by Susan Calvin to a reporter, in the first story we learn about a tender relationship a girl has with her non-speaking robot nanny. The nanny robot is one of the earliest robots made by U.S. Robots Inc. Here, feeling increasingly uneasy about her daughter's intense relationship with her robot, a mother sends the robot away. But because the girl wouldn't give up her love and efforts to re-find her best-friend robot, her father eventually decides to set up a reunion and returns the loving robot back to his daughter. So here we have a nice, innocent and sweet story that illustrates the 3 Laws acted out by an early-model robot.

Unfortunately, as the chapters move on and as the robot technology more and more increases over time, U.S. Robots Inc. has a harder and harder time controlling the behavior of their robots even though the 3 Laws engineered deep inside their positronic brains still and always remains true. Asimov sets up tales where the engineers have an increasingly complex time rectifying the problems precisely due to the robots acting out the 3 Laws. Each story builds upon the last in terms of the engineers' and Susan's increased needed ingenuity to solve the foregoing problems with the more advanced robots.

Things finally come to head in Chapter 9 called "Evidence."

**Please keep the following in mind:*

When Susan tells the reporter the following story the protagonist is dead having atomized himself, somewhat akin to modern-day cremation though in this futuristic case, leaving no DNA behind.

It involves a man named Stephen Byerly, 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 Byerly is indeed a robot. Countering this claim and thus what eventually gets him elected as mayor, Mr. Byerly punches a spectator at one of his speeches who is egging him on to hit him thus proving that Byerly is not a robot. If he hits this man, the 1st Law of Robots would be nullified. Again, it is:

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

Actually, My. Byerly is a robot and the man who created him also created the robot who posed as a man at the riotous speech. So in fact, Byerly never broke Rule #1. Stephen Bylerly 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 of the book, "The Evitable Conflict."

At this time, Byerly is still the world coordinator. Earth is now fractioned off 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 advanced and extremely complex robotic brain entirely dictated by positronic circuitry. Susan Calvin explains to Byerly:

> "Stephen. 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.'"

Things at this time have become askew and that is why Byerly summons Susan to his office. For the first time ever under the Machines' precise planning schedules for societal balance, there are economic allocations that are relatively inaccurate. For Byerly, this causes great concern and he believes the 'Society for Humanity' —a vociferous anti-Robot group—is behind all of this. But Susan sets him straight. The robots are satisfying the wants and desires that people may not themselves be aware of. Still the Machine's actions might hurt some individuals, for example temporary unemployment, but for the greater good of what people mostly unconsciously desire the Machine is acting in accordance.

The brilliant 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!"

And with this the fire in Byerly's study flames out.

WHY SUMMARIZE *I, ROBOT*

Artists foretell the future. In society, they harbor the creative genes. Their creative output pushes science forward with the blueprint to what science can shoot for. For both good and for bad. On the original Star Trek, there were hand-held devices that crew members used to talk to each other where they could see the person they were speaking to. And 50 years later, science made it happen—Apple's FaceTime. There are endless examples where artistic imagination turns into science's goalposts.

The specific **I, Robot** edition that I used to write this part of the paper was published by Fawcett Crest Book in 1970. Its anonymous "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 bold type above is commonplace with how people speak about good science fiction. Saying this is certainly not out of line. Allegorically, it matches what many peoples' sentiments are with regards to the current AI atmosphere.

AI is a double-edged sword. A very powerful double-edged sword! No need here to replicate what we have extensively said elsewhere. But what I would like to point out from Asimov's book is that THE THREE LAWS OF ROBOTICS was explicitly ingrained deep inside all the robots and the Machines to ensure man's protection from his creation. In this way, a very stringent, unconscious Freudian Super-ego was established. Notwithstanding all of this, still in the end the robots became the masters of man's domain.

I am reminded of something I read by Freud, which at the moment I don't remember where, where he stated something along the lines that science has taken over the role that religion once had in earlier days, but for good or for bad, science has REAL WORLD EFFECTS.