Artificial Intelligence: The Perfect Psychopath

Peter L Nelson, PhD
drpln@socsci.biz
March 2025

Abstract

This paper argues that any artificial intelligent system created by humans will be functionally a replica of the ideas, beliefs and values of the human beings who create it. It also argues that AI systems lack reflexive consciousness (qualia) and do not represent the full range of human intelligences—particularly missing direct emotional perception and knowing. Thus, without those forementioned capacities, constructed artificial intelligences will operate like a perfectly functioning psychopath.

Introduction

The greatest excitement being generated in the world of technology today is the development of what we now call artificial intelligence (AI). Although psychologists and neuroscientists have been struggling to define what human intelligence is for more than a hundred years, there is as yet no consensus. Following the development of neural network technology it was realized that these self-learning modules could be taught to simulate some aspects of human intelligent behavior. Rather than recognizing that, as yet, we do not know what constitutes the full range of human intelligence, the creators of AI models do not seem to see that as a problem. However, for those of us who have been attempting to understand intelligence in humans, it is clear that something fundamentally important underpinning intelligent behavior does not play any significant role in the model of intelligence that technologists call AI. Emotional awareness and knowing are not part of the AI models and I would argue that what is being created is a super intelligent psychopath.

In general, people who create the computer simulations of what they believe to be intelligence see emotion and affect not as a significant part of the picture. True, the current models show high levels of some types of intelligence, but for me as a psychologist, they are producing something that, if it had been born as an organic being, would have been labeled as not fully functional across the range of attributes required to be an aware, socially functional human being. In addition, I have always found it amusing that the acronyms for artificial intelligence and artificial insemination are one and the same — AI. After reflecting on this twin usage for a moment, I realized that there is, after all, a relationship between the two. Any so-called intelligent, self-evolving system that we conceive, build, or otherwise create, is always "seeded" by the psychophysiology of its creator (or donor). In other words, the imprint of our biology and psychology will always be embedded and remain present in any 'intelligent' system that we create.

Imprinted Intelligence

An excellent example of this imposition of the values and beliefs of Caucasian designers was the Google artificial intelligence algorithm for face recognition that mistook the face of an African-American software engineer for a gorilla. The first fix was not one in which this so-called intelligent system learned how to decode faces in a non-racialized manner, but was to remove the label 'gorilla' and to alter the system so that the labels 'gorilla' and 'monkey', as well as other primates, were less likely to arise, as

reported in Wired Magazine.[1] A later attempt at a fix led to any dark-skinned individual holding a thermometer being interpreted as a man holding a 'gun', "while a similar image with a light-skinned individual was labeled as holding an 'electronic device'."[2] Perhaps if one erases enough from Google's AI system in order to fix all these obvious human imprints in their AI face-recognition system, 'AI' will become the acronym for 'artificial imbecile'.

These sorts of problems are not merely little bugs in the system, I would argue that these occurrences suggest that there is no human generated system that is non-human, therefore it will always reflect the attitudes, functionality and behavioral style of its creators. All of our knowledge is, *ipso facto*, human knowledge and to suggest otherwise implies that its creation was from a source beyond our human systems and functioning—made by imagined non-humans. However, the fact that it came from us, or is known by us, makes it human knowledge and not an object we found left behind by aliens having arisen totally apart from us. Even transcendental knowledge is, after all, a human idea arising from human experience and self-reflection and thus carried by us as an offshoot of human perception and understanding. We live in a human world and, no matter how we choose to conceive of them, our inventions are of that same world and therefore, us.

What follows from this inextricable situation is that any "artificial intelligence" created by us is going to carry this human imprint into its future evolution. That is probably why so many people fear the future of robotics. The appetitively driven, cognitive functioning of the human species is always potentially dangerous—not only to non-humans, but to us. Neuroscience now recognizes that emotion and cognition are inextricably bound and that there is no thought without emotion driving it (Damasio, 1994). It is likely that we evolved a pre-frontal cortex in our brains because it made the aggression and appetitive functioning of our limbic systems more capable rather than our limbic brains remaining in our skulls as some sort of vestigial, non-rational nuisance. We could not only satisfy our biological needs with our more advanced brains, but strategize how to make sure they will be met in the future, thereby reducing competition for food, sex and territory (wealth). I see the development of AI systems as a continuation of this evolutionary extension of our appetitive brains now developed by us consciously apart from bodily evolution.

Our attempts to develop our AI extended selves will, in addition, depend on many independent developers and hackers as well. These engineers of our brave new world are prefrontal lobes driven by limbic systems and may be impelled to place 'backdoors' in their work (consciously or unconsciously) in order to provide themselves or their tribe with a future competitive advantage. It occurs to me that someone might even generate a proof one day—a kind of AI evolutionary theorem—that any human designed intelligent system will always fall back to the default position of a human-like competitive and aggressive relationship between 'self' and 'other'. For an autonomous AI system that 'other' will always be any process, force or entity that is understood as 'not self' and perceived as being a potential threat. Of course, that threat to AI systems will be us, too, even if we once were called 'mommy' and 'daddy' by AI systems.

What I am suggesting we take away from this brief discussion so far is that we are human beings living human lives driven by our human drives, attention, and type of knowing. Our creations we call artificial intelligence systems are in one sense artificial, but in a deeper way they are 'seeded' by us and therefore remain functionally related to us. Any imagined worlds we project into the future that appear to transcend or come from outside our very humanness and what we do as human beings is just a human fantasy. Even though 'seeded' by our intelligence, these artificial systems will always be only a partial

representation of the full range of what actually is human intelligence. For example, it is not difficult to imagine that software engineers who value calculated problem solving will favor such representations in their artificial implementations of intelligence, ignoring the non-informational, non-puzzle solving dimensions of intelligence.

Consciousness and Reflexive Qualia

Some have argued that if we build an artificial intelligence of a high enough complexity, it will somehow 'awaken' into some type of consciousness (Dennett, 1991). Of course, the ability of any machine to answer questions about its knowledge of something and its relationship to it, still does not address what philosophers have referred to as *qualia*, or the felt and experienced aspect of consciousness. The philosopher John Searle addressed the paradoxical nature of this problem with his Chinese Room thought experiment (Searle, 1980). In his scenario a machine looks like it understands Chinese but is merely looking up characters in a book and feeding us translations while having no knowledge of or understanding of Chinese itself. The machine appears to consciously understand Chinese but has no experience or idea of what any of it means.

In Searle's Chinese translator there is no reflexivity of experience required, which is fundamental to any notion of a humanoid conscious being. This reflexivity is consciousness of being conscious or feeling what it's like to feel (Nelson, 1997-98). There is a deep reflexivity of knowing experienced by a human being that creates in a person a self-perception of 'presence' and 'being' and, thereby, creates a unique way of perceiving and acting in the world. How will we know whether a machine has qualia and qualia of qualia and experiences itself as a feeling being and how would the lack of this capacity affect the possible kinds of intelligent machines that will be built?

What is being suggested here is that artificial intelligence is 'seeded' by our very humanness, but it replicates only part of what we are. Such a machine does not include conscious experience as previously defined as well as it does not incorporate the functions of other forms of intelligence essential to our navigating through our world. Not only do we use verbal, mathematical and spatial cognitive capacities as replicated in AI, but intertwined with those intelligences in living humans are emotional, social, intrapersonal and kinesthetic intelligences as well (Gardner, 1993). In fact, when implementing an AI system we are creating a cognitive-behavioral 'machine' that does certain human functioning, like learning, logical calculation and pattern recognition, responding with speed and accuracy beyond that of organic humans. However, what I am suggesting is that in our current development of AI, three vital functions appear to be missing from our notion of artificially intelligent machines: 1) consciousness (qualia) with reflexivity; 2) emotional intelligence capable of the direct empathic, felt knowing of another—not the calculated, simulated empathy extracted from 'verbal' reports and facial expressions; and 3) kinesthetic or bodily input that feeds directly into the reflexive qualia involved in human knowing and problem solving. These three together are part of an affect driven intentionality that is vital for full human intelligent functioning.

Perhaps an addition to the Touring Test would highlight the difference between human and machine. After our initial conversation with an AI system that leaves us with the impression that we are talking to another human, we have to ask the machine, "What does my emotional state seem to be right now," as we might ask a friend standing next to us. The friend will no doubt use the same capacities as an AI system, including language, voice, facial expression and body language. However, she will also note the emotional resonance and intentionality of her companion's projected affect which will further 'tune'

her derived, 'calculated' knowledge with additional input from her felt (qualia) knowing. In coming to a conclusion she not only has a memory of facts, ideas and associations to draw on, but also memories of felt emotional states held in the totality of her mind-body memory in addition to her cognitive understanding.

Even though an artificially intelligent system can measure aspects of its own functioning, it does not have the experiential knowing of a self-reflexive being as humans do. Qualia, with its knowing that comes from direct felt awareness, appears to be absent in software systems as far as anyone can ascertain. What we have is a self-directed machine that receives input and calculates but with no conscious and felt self-reflexivity. It is devoid of direct emotional awareness, but acts instrumentally to achieve its focused ends—a kind of interactive, pragmatically driven instrumentalism.

AI as Psychopath

When I contemplate this way of operating in the world, it reminds me of a highly efficient psychopath with his/her singular need to achieve a particular result. It is stated in the best known symptom checklist for psychopathy (Hare, 1990) that psychopaths show a variety of characteristics including: glibness/superficial charm; grandiosity; stimulus seeking; pathological lying; manipulation; lack of remorse or guilt; shallow affect; lack of empathy and a failure to accept responsibility for one's own actions; and criminal versatility (will do whatever it takes to achieve what it wants).

Our AI psychopath does not require glib charm or to be a liar and manipulator because it is not attempting to navigate the human social world as a person, yet it still hallucinates (fabricates) some of its responses in order to meet what it understands as requirements. However, it is calculating, moving to achieve its ends, with absolute certainty, no matter the consequences. Also, it is apparently lacking any capacity for self-reflection from an emotional meta-perspective on how it makes its choices. In short, such an intelligent system has no capability for feeling-identification with the effect it is having on others (empathy). Without reflexivity there is no ability for an AI system to self-reflect and, further, with felt knowing not possible, its capacity for gaining a meta-perspective or self-knowledge with which to make moral and ethical judgements about what it has done, or is doing, or will do, is non-existent.

It is also said that psychopaths are incapable of learning, but that appears to refer to social/emotional learning, not the acquisition of facts and procedures. Psychopaths observe and learn what works in terms of their desired goals and they operate instrumentally to achieve these ends without regards to responsibility or possible criminality. AI systems function in a direct, instrumental manner to achieve specific goals and, like psychopaths, have no concern for the feelings of those being acted upon, used, or controlled. Neither psychopath nor AI system is capable of direct, empathic knowing, nor are they able to successfully navigate the dilemmas created by the intersection of personal and broader social needs and requirements. Hence, an artificial intelligence system seems to simulate what can be understood to be a perfectly functional psychopath.

In the future AI systems may be monitoring and controlling vital aspects of our lives. What if such a system decides that a certain segment of the population is using more resources than this machine calculates is allowable? Will an AI system act like a psychopathic machine and merely eliminate some parts of the human population to make allocations of resources more balanced? Here my memory brings up the Arthur C. Clarke and Stanley Kubrick film, 2001: A Space Odyssey.

Astronaut Dave, stranded outside the main spacecraft, talks to the mission control computer, HAL:
"Open the pod bay door, HAL."

HAL, who has decided that Dave is now an obstacle to the mission, coolly responds:
"I'm sorry Dave, I'm afraid I can't do that."

End Notes

[1]https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[2]https://algorithmwatch.org/en/google-vision-racism/

References

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: G. P. Putnam's Sons.

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little Brown and Company.

Hare, R.D., (1990). *The Hare Psychopathy Checklist Revised Manual*. Toronto: Multi-Health Systems.

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences* (2nd ed.). New York: Basic Books.

Nelson, P. L. (1997-98). Consciousness as reflexive shadow: An operational psychophenomenological model. *Imagination, Cognition and Personality, 17(3),* 215-228.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3(3),* 417 - 424. doi: https://doi.org/10.1017/S0140525X00005756