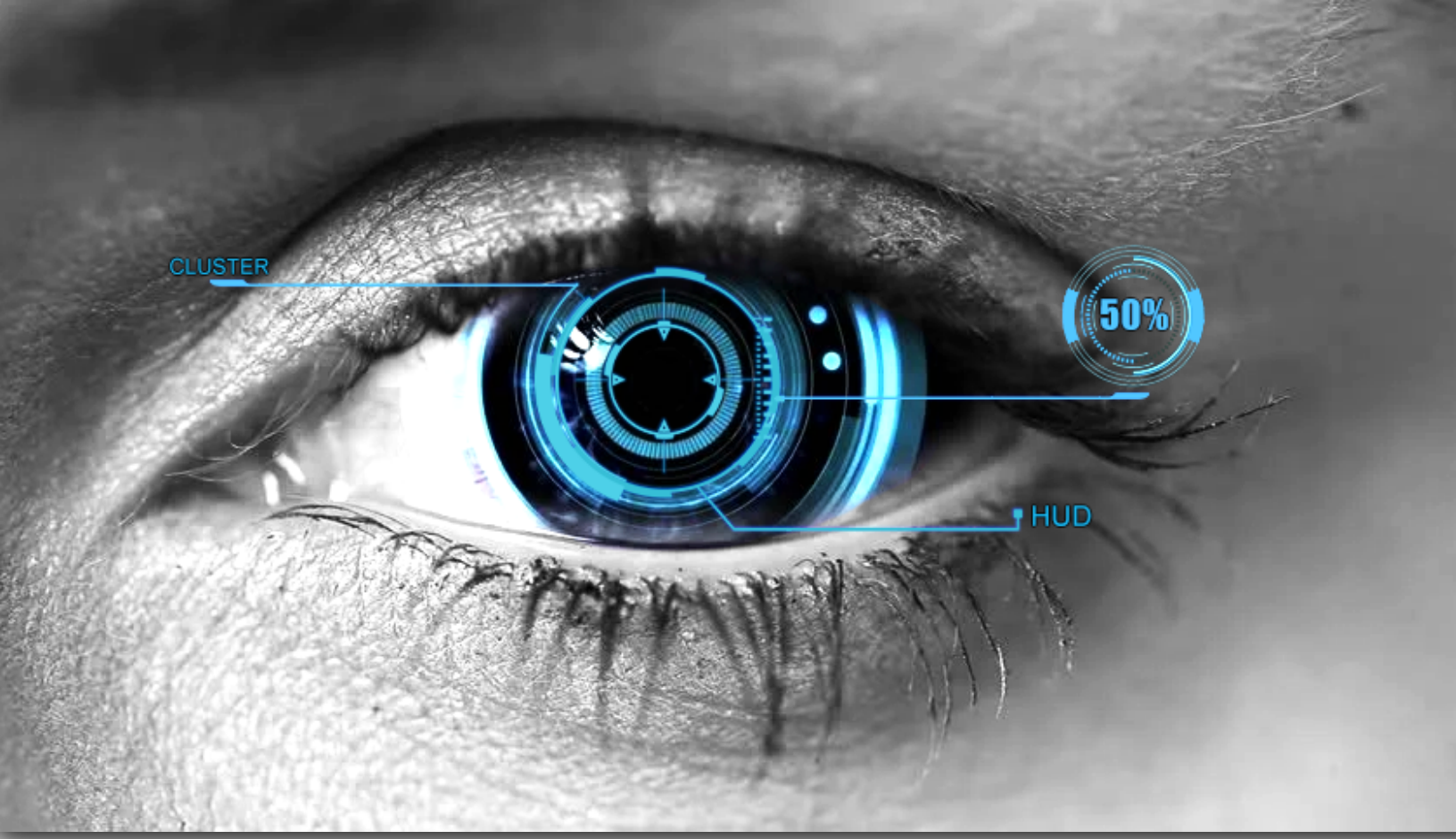
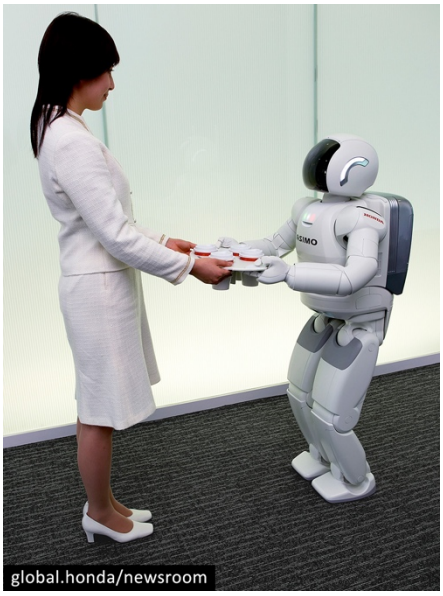


The Technological Risks of Sentient AI

by Chad Jordan - July 2nd 2022

We all know Artificial Intelligence goes back decades in research and testing. Even with the first official recording in the early 1950s with Christopher Strachey and his Checkers game, there's been an ongoing desire to make machines think. In the coming decades most of these applications have been in very helpful areas of banking with omnichannel banking, marketing with automated decisions based on data collection, and entertainment with post-production in film editing, user experience, storyboarding, and advertising. If you were born after 1980, you most likely associate AI with video games. During the early 80s, it was Honda that researched and began manufacturing their first AI robot ASIMO ***(Advanced Step In Innovative Mobility)***, later revealing it in the year 2000. AI has been implemented in many practical and entertainment applications over the years, but as technology advances, so do our capabilities to test new areas. In my personal experiences with AI, I've programmed the Arduino microcontroller using C programming to move robots from point A to point B through a maze, but I've mostly dabbled with AI for game development purposes. These practices are meant more for experimental purposes, not necessarily designed to explore what I call, *uncharted waters*. It's been said that with great power, comes great responsibility, so from an ethical standpoint, the question is, how far do we push the boundaries of AI? How far is too far?



For example, let's look at the ASIMO robot from Honda. When artificial intelligence is created to help others in daily tasks, this is a practical application that makes sense to have this form of machine assisting those that need help. We conclude this because we know the ASIMO robot has only been created to communicate and assist with day-to-day tasks. It doesn't question said tasks or their existence. It understands essential communicative responses but holds firm to finite thinking. If a human requests an action from ASIMO, they ask if a given task can be carried out, and ASIMO's response is, "Yes, I can." It understands that a request has been given and that it needs to carry out the task until it's properly completed. Natural language processing is a fantastic technology, and while not required in my specific track of study in computer science, I went to school with colleagues at my university who studied it in great depth. In 2002 one of my roommates was a double major in computer science and psychology. He was determined and a little obsessed to learn more about how we could advance beyond language structures with textural and vocal tones with computers. He explained how he was working on a program that unites cognitive human neural science, with machines to learn and develop emotional responses. I remember turning to him and with a

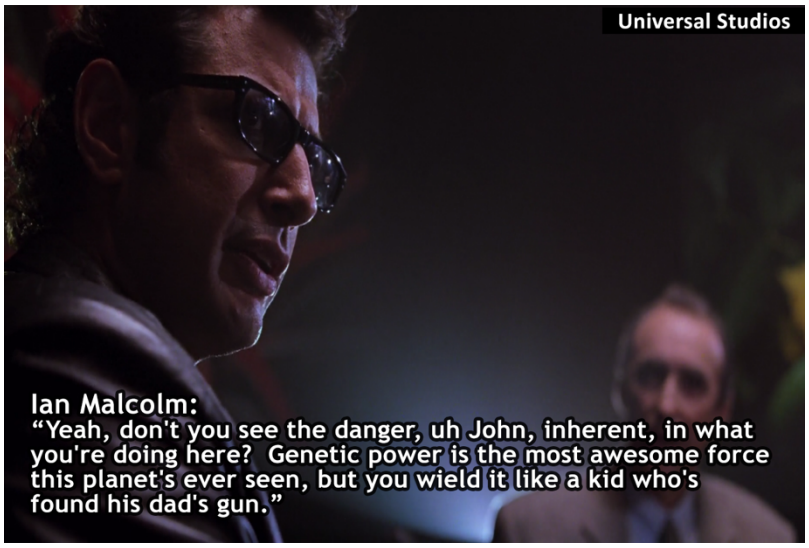
slow, trepidatious response of, "…Why?" His response was in lieu of curiosity, technological exploration, and being one of the first to successfully implement it neither of which I felt were good reasons to toy with something of that magnitude.

As humans, this is what we do. We allow curiosity and the notion of the, *"Because we can"* mentality to overthrow the morals and ethics of computer technology. The idea of computers with emotions is nothing new. For years we've seen how it becomes abused in films such as *Blade Runner, Wargames, The Animatrix, Artificial Intelligence, iRobot,* and so on. Are these films really that far off from the possibilities that could occur in reality if we advance too far? If you consider that we do successfully keep pursuing these endeavors, what happens when something goes wrong? This has always been the case even with *"the perfectly written system"* Programming has always followed Murphey's Law, "What can go wrong, will go wrong" it's just the nature of technological imperfections. If you're faced with a computer/machine that responds emotionally, it is inevitable that it will go wrong. It will happen, so what do you do when it finally does, and also, what does that look like? How do you prepare for that? If you're faced with a machine that develops its own independent thought from your directives what will you do? Depending on the severity of the situation, you have to have the means to either contain it or destroy it, and by doing so, you've wasted a great deal of time and resources. Not to mention, your potential safety and the safety of others in more complicated situations. For a number of years, AI has presented growing concerns related to **privacy** with smartphones. We all know those ads are not just conveniently showing up two minutes after some specific keywords are mentioned in our conversations. This is one targeting method known as **internet advertising**. Artificial intelligence is increasingly being used to craft advertisements for internet users. With these types of practices in place, we should be concerned about how some users abuse the system by getting their hands on other people's personal data. We all know how harmful this can be for everyday internet users. The same is said for AI **surveillance** through self-learning in airports, large-scale cities, and space. This began as the computer learning how to reason by scanning objects of interest within a scalable viewpoint. These self-learning capabilities are continuing to branch out and now it detects objects in the real world. One of the big problems with AI learning these types of capabilities is when it moves past the intended scope and creates unnecessary identity issues in security. This way the AI is constantly learning and improving in ways that can be harmful to us.

Another area of concern that really hits close to home for me is the **environmental effects** that AI has on the world. According to Europa the damage from fuel emissions is considerable. AI has obviously been around for nearly half a century so everyone knows it is nothing new, but training a large AI model to handle human language can lead to emissions of nearly 300,000

kilograms of carbon dioxide equivalent to about five times the emissions of the average car in the US, including its manufacturer. AI is one of the technologies that is supposed to help us with climate change with early predictions not create more problems with emissions. Something we've all had on our minds for many years, especially only several years back between 2017 to 2018 is the rising threat of nuclear warfare. This growing concern is **malicious use on a catastrophic level**. Nuclear warfare is by no means an exaggeration or too far of a stretch. In every society, fear is what fuels the need for technological advancement. As a human race, we're terrified of the unknown, and to that end, we're even more terrified of being unprepared.

While there is plenty of other concerns with AI, another area that is becoming far more apparent these days is **AI Consciousness**. How does a technology become self-aware of its own existence and volition? We've simply concluded for many years that this has always been science fiction, and will never come to fruition. The majority of programs that I've written in the past have mostly remained static. However, when programming AI we know it's never designed to remain static, change is inevitable. The algorithms that are being used when writing the program should spawn change in directives even if only a little. Some of the most common functional languages that are found to be used a lot for writing AI are Haskell, Lisp, and Smalltalk. Some of the growing concerns that I've mentioned above have been occurring for a long while at this stage, but some of my deepest concerns at the moment are closely related to sentient AI. We have several moral dilemmas that we face when looking at this subject. Quite possibly one of the most important questions to ask ourselves is, what are we wanting to accomplish with this area of study and why? We know what ASIMO does when carrying out tasks for humans, but why does AI have to be self-aware or have the ability to decide its own will? How about the inherent danger involved in this? Anyone who has seen the 1993 Steven Spielberg film, *Jurassic Park*, knows that one of the most important scenes of the film is when they are all sitting down to lunch and having their discussion regarding why John's decision to exhibit sixty-five million years of prehistoric evolution is a bad idea. The entire conversation from that scene actually overlaps quite a lot with the dangers of experimenting with sentient AI. We have technological advancements right in front of us and *"we're so preoccupied with what we're doing that we haven't stopped to think if we should."* As a human species, we're terrified of the unknown, yet also terrified of not knowing. While I can appreciate technological exploration and discovery, the simple fact is some things were not meant to be explored. Yet, here we are using the tools



of science and technology to push the envelope as we perpetually do to go further or to be known as the first to accomplish it. The problem is, that if we keep looking deeper, we may uncover something we don't want to see, and uncovering that is the true terror of discovery. This terror is known as **AGI and Superintelligence**. AGI *(Artificial General Intelligence)* is when the AI has reached the human level of understanding. At this stage, the AI can handle any task that any human and the average human can do on its own with no assistance. If or when AGI surpasses human intelligence, then superintelligence is reached. Superintelligence would be vastly superior to humans and therefore have no further need for mankind anymore. What use are we if there's another entity smarter than us, and doesn't require us to maintain it, or build new entities because it can handle it on its own? AI technology is accelerating at an alarming rate and global corporations and governments are racing to claim the power of AI as their own.

This brings me to some information in recent news from the Bloomberg post regarding AI issues at Google. Google has had controversial issues with AI for years, so it's nothing new, but more recently from June 14th *(only 3 weeks ago),* this article states that a software engineer who has now been suspended, made public comments claiming *one of Google's chatbots to be sentient in nature*. According to Google and the AI community, the software engineer's comments had no basis on fact. Regardless if this is the truth or is otherwise merely being covered up, the article conveys concerns *"if it's possible that AI can engender real-world harm and prejudice, whether actual humans are exploited in the training of AI, and how the major technology companies act as gatekeepers of the development of the tech."* These are absolutely viable things to consider and I especially do not doubt the probability that training in AI is exploited, and global corporations act as gatekeepers for maintaining the enhancement of AI. The more hardware that is provided to AI, the faster, and smarter it becomes. Larger corporations have plenty of money, power, and resources to sway the continuing maturity of advanced AI systems and how they are integrated. These are the things that are beyond any of our control. If AI is going to continue to evolve even in sentient ability, it's going to happen whether we want it to or not. There's just too much power beyond us that makes these decisions. I think part of the concerns related to this article is how some of these engineers are trained to remain silent with AI practices. I personally know two engineers that have worked at Google as developers, and much of the intelligent systems projects they were doing was unable to be discussed. From a legal standpoint, I can identify and respect confidentiality especially if there is something a company is wanting to have patented for protecting their own work or if it's a government contract. However, if the truth is that there are ethical issues related to the silence because of questionable practices, this is where they should reconsider the ***type of work*** they are performing. This goes back to the implementation of the chatbots. The article mentions, *"Chatbots are programmed with trillions of words from the internet in order to mimic human conversation. The conclusion from Google's engineer was that the AI was a sentient being that should have its own rights."* He said "*the feeling was not scientific, but religious. Who am I to tell God where he can and can't put souls?"* he said on Twitter.

In my own personal experience, I had a somewhat similar instance with a chatbot back when I still used Amazon. Initially, I thought it was an automated system that would simply route me with a few responses to the accurate representative that I needed. Throughout the

conversation it was actually portraying very human-like responses that could be classified as emotional, and I found myself asking it, “Is this an automated system, or am I actually speaking to a real person?” The chatbot responded, “yes sir, I’m real.” After several more minutes, I noticed it was taking me in circles, and starting to give almost the exact same responses as earlier in the conversation, and I decided to test it by asking what its favorite color was two different times. When I received two different colors each time I asked, I realized it was most definitely not a real person but in fact, an AI that Amazon was still in need of improving. When I got off the conversation I was very much in a state of discomfort. I told myself that the conversation was way too real to be some generic AI, but I knew with the repeated questions and the way we were going in circles, there was no way it was a person I was talking to. The weird thing is, Amazon gave them names, like Paul, and Ashley. The one I spoke to was Alex, and it’s so unsettling that Amazon would give them human names when they were definitely not real people. The AI was impressive enough that it had me going, but failed when it began to repeat itself, and provided me with two different colors from two different instances during the conversation. If it’s a chatbot, then call it “Chatbot” or “Chatbot_5” or “Chatbot_6” don’t give it a human name like Josh, or Kim. The whole experience was far too bizarre and I never chose the chatbot option again after that, plus I never did reach my representative so the entire exercise was futile to begin with.

There are a lot more issues regarding AI compared to the small amount that I have covered in this article, and all of these various aspects of AI should really be concerning us. As humans with free will, we can still pick and choose what companies and products we want to use or support. E.g. I have canceled my account with Amazon for numerous reasons and haven’t used them in more than 3 years and I couldn’t be happier. Aside from Amazon’s reputation with how they treat their employees, their integration of AI practices is very questionable with these chatbots that they use. The same concern should be coming up with regard to Facebook/Meta. During 2009, I was on Facebook for about 8 months before I realized how bad it was for me and others around me. I’m not just referring to the obsessive addiction of checking accounts too often, but rather the integration of AI that has been used to pull people’s data. Again, global corporations with power like Amazon, Meta, and Google are the kind of companies that we should really be questioning their methodologies with their implementation of AI and data retrieval for improper and unethical uses. These organizations have a responsibility to uphold but since we can’t rely on them to do that, we can make the decision if we want to support them or not.

It should go without saying that just because we may have the capability to explore something, doesn’t mean that we should. I’ve held firm in my beliefs with this from the beginning and it won’t ever change. Technology exists as it does, in a finite, emotionless state for good reasons, and we should never toy with that. Part of being human is having emotions regardless of the irrationality and illogical nature. As humans, we were created to feel because even emotional thought still helps us to distinguish between choices that are right and wrong. The problem with emotional thought is that it’s still irrational in nature. The reality is that we try to find affirmation and validity in our work. We reassure ourselves and others that we’re changing the world for the better, but we can’t create the perfect system because perfection is unknowable.

Therefore, we're always striving for more. Many companies push the ideology of "never settle" but the problem with that mentality is that it teaches us that we're never really good enough where we are, and we must continue to strive for going bigger, faster, and stronger that we're not even aware of when or how to stop when we should. Sometimes, enough is enough and we need to back away from the keyboard and think about what we're doing.

Resources Used:

- How Stuff Works
- Honda Global
- Getty Images
- Universal Studios
- Europa
- Bloomberg