Proposal, UTC2113 Gaming Life

Rubesh Suresh

October 13, 2023

Abstract

In this proposal, I elaborate the parameters of my work, and certain theoretical underpinnings, motivations, and plans pertaining to that. I explore the interface between artificial intelligence, consciousness, autonomy and pain.

The Opening Act

"I control a slave, a dog, a worker; but if I establish complete control somehow, as by implanting electrodes in the brain, then my subject is little more than a tape recorder, a camera, a robot. You don't control a tape recorder—you use it." (Burroughs et al., 1999).

In an era where technology burgeons at a pace once deemed unimaginable, we find ourselves at the crossroads of ethical dilemmas, philosophical contemplation, amidst ground-breaking innovation. As Burroughs suggested, there exists a fine line betweeen control and use - The former implies a certain degree of autonomy, while the latter implies a certain degree of subjugation. What happens when this line blurs, and Artificial Intelligence (AI) starts mirroring human-like consciousness? What happens if in the near future, we have to *control* machines instead of simply *using* them like we always have? This proposal delves into the fascinating, albeit complex, interplay between AI and human existence, the interface between artificial intelligence, consciousness, autonomy, existence, and pain. By drawing parallels from existential and ontological philosophy, I venture into a hypothetical realm where machines could potentially transcend their electron-based confines, reaching a state that mirrors human consciousness. In my work, with the aid of an inquisitive and unorthodox application, I attempt to simulate this potential reality, offering a thought-provoking look into the (potential) future of human-machine interaction.

The Concepts

Thus far, our usage of computers has been unidirectional. We subject our utmost wishes and desires on these machines and when a computer, despite its being run on millions upon millions of lines of code with an unfathomable number of functional intricacies, slows down for just a moment revealing its flaws we get angry and dissatisfied, perhaps

even discarding it eventually. As such is the fate of the computer, despite its billions of transistors and countless offerings, it must live up to the standards we expect of it lest it be discarded.

Absent are the moral qualms presented in Burrough's quote - a computer is a non-conscious machine, designed purely to exectue instructions step by step, and at the binary level this is merely a sequence of 1s and 0s. Even lower, at the atomic level, this is merely the flow of electrons in silicon transistors.

The computer is neither a slave, nor a dog, nor a worker - it is a tape recorder (literally, too). It is a tool, a means to an end, a medium for our desires. It is not a being, it is not conscious, it is not alive. It is a machine. We are inclined to differentiate ourselves from machines, because we think, we feel pain, we are conscious, we are alive.

Unfairly, we define machines at the atomic level and ourselves at the highest level of abstraction. However, the constitution of the Human is not too different from that of a Computer. The human experience can also be reduced to the flow of electric impulses through the nervous system, through neurons in the brain, to neurochemicals such as dopamine, serotonin and oxytocin that give rise to the complex emotions we feel.

These neurochemicals themselves do not have emotions, and neither do electrons. So what then can explain consciousness and our conscious experience? Clearly there is a stark contrast between our own consciousness and "the consciousness" of a machine, despite there being minimal, or dare I say, no difference at the lowest level. Somewhere between the levels of these base particles at the lowest level of molecular composition and the pain or pleasure we consciously perceive at the highest level of our conscious experience, somewhere in this midst there must be a spontaneous coalescence of these particles to form something greater, one which as a whole exceeds the sum of its parts, an explanation that seems to entirely contradict our modern understanding of physics.

"Property dualism postulates an ontological distinction between the attributes of mind and matter. It contends that consciousness cannot be reduced to explanations grounded in neurobiological processes or the axioms of physics. Essentially, when matter is organized in the appropriate way, mental properties emerge". (Searle, 2002). If we are more than the atoms that constitute us, is it possible to then say, that machines could become more than the electrons that constitute them? Is it possible for the so-called "mental properties" to emerge spontaneously, and in doing so, becoming a whole larger than the sum of its parts? Within the scope of my work, I aim to bring this contemplation to the forefront of our awareness. I attempt to explore an alternate reality, one which might even be of a future yet to come, where hyper-evolved computers - or AI *Entities* - redefine the human-machine interaction paradigm. As we become exponentially reliant on technology, I want to explore what a reality where the tools we now take for granted turn against us, one where the emergence of consciousness fundamentally alters the in-

teraction paradigm.

The Work

The synposis of my work is as follows: I create an application (the bot) to represent the AI Entity in question. The (human) user can type commands to the bot to which the bot will respond accordingly. The bot resembles a simple task manager and the user will be able to functionally utilize the bot, such as adding and deleting tasks by typing the associated commands. Most importantly, the user will be given an additional choice to send negative "hate" messages to the bot. The bot will be able to understand that the user is doing so, and will respond with a message that indicates it is in pain. The bot will be programmed to remember the pain it feels, with an "I_FEEL_PAIN" pain variable. Additionally, every time the bot is run, it will be programmed with a maximum pain threshold, "$I\_CANT\_TAKE\_IT\_ANYMORE$", determined probabilisitically at run-time, which is known neither to the user, nor myself, the developer. The user's treatment of the bot will have one of two consequences.

In the first variant of the bot (BV-1), it will be programmed to direct any pain it feels towards itself. The outlet of pain for BV-1 would be its own body and composition, which translates to its own digital source code. As the user continously mistreats BV-1 and its pain variable increases, BV-1 would internalize this pain by slowly devouring its own source code. This process would be made tangible to the user as BV-1 would lose more and more of its functionality (e.g. the user might not be able to delete tasks anymore) and once BV-1 loses a functionality, the source code pertaining to that functionality is permanently erased from the computer. Eventually, when the pain variable exceeds the pain threshold of BV-1, it will delete its entire source code, and the user will not be able to start the application again. BV-1 would have, in a way, committed suicide. The only way for the user to recover from this is restart the whole demonstration, as the application and all the data contained (about the user's tasks) would be corrupted and cannot be run.

In the second variant of the bot (BV-2), it will be programmed to direct any pain it feels outwards, against the "world". In the context of BV-2, the "world" refers to its own digital confines, the host computer that it is run on. The user's messages will have a different effect, that is, as the user continues to send hate messages, the bot will obfuscate its functionality and refuse to perform certain tasks. Similar to the first scenario, this threshold once reached, is irreversible, at least directly. However, looking at the source code, unlike the first scenario, the code will still be present, symbolically indicating that BV-2, while it is still able to perform the given task, is actively refusing to do so. With some minor technical know-how, the user will be able to circumvent the BV-2's refusal by bypassing the application itself and modifying the source code manually, and this is intentionally the case in order to fundamentally differentiate BV-2 which harms the world from BV-1 which harms itself.

The key difference between BV-1 and BV-2 is that once the pain threshold is reached, as opposed to "taking its own life", BV-2 will direct its pain upon the world (again, referring to the computer it resides in) and in technical terms this translates to destroying the computer with a series of harmful and malicious shell commands*. The term harmful here refers both to the user and BV-2 – The user's computer would essentially be ruined, and ironically, BV-2 itself will end up "dying" as it "sets fire" to the world that it needs to survive.

The probabilisitic nature of the pain threshold emulates human emotional unpredictability, and the user will not be able to know the pain threshold of the bot until it is reached, but by then it would be too late. The user, also being an autonomous agent, will be able to make the choice whether they want to take the risk and subject the bot to pain by mistreating it. Both variants will be made available to the user in my work, and the user will be able to choose which variant to interact with, one at a time.

Parameters and Limitations

In the exploration and development of my work, I aspire to endow a digital Entity with the capacity for autonomous choice. It is crucial that this self-directed capability is then paired with the ability to experience and process pain. My intention is not just for the Entity to recognize this discomfort internally, but to externalize it in a manner that's palpable to the user. By rendering the pain tangible and projecting it outwardly, I aim to foster a deeper comprehension in the user — a realization that the Entity isn't just a mere sequence of algorithms (even though it is) but possesses a semblance of consciousness. The AI's interaction with pain is designed to mirror the human experience, providing an avenue for users to better empathize with and understand the bot's semblance of sentience.

However, it is imperative to note that within the confines of this work, the complete breadth of connections between every facet cannot be exhaustively explored. In my work, I make certain assumptions such as the link between consciousness and pain, the rigourous exploration of which is outside the scope of this work. I predicate the basis of my work on a general understanding of these various concepts and intend to focus more on the development and exploration of human-machine interaction.

Additionally, In my endeavour, I have to acknowledge that the bot, it is at its core, still a computer program, written in computer code, and executed by computer chips. It is not a sentient being, it is not a conscious entity, and it is not alive. It is merely a simulation of one, and as such, it is limited by the constraints of the medium it is presented in. That is to say, that all "sentient" behaviour it exhibits, all the ideals of "autonomy" that underlies its existence, its mimicry of consciousness is all fundamentally programmatic in nature. Thus, all its "expressions" of pain are the bot merely executing a sequence

of clever* programming tricks, intended to simulate the behaviour of a sentient being, sufficiently enough to convince a human user that they are interacting with one.

In order to better demonstrate my work, I segregate the two types of pain projection into two distinct variants, however, in reality, the projection of pain is not always so clear-cut. In reality, humans often project pain both inwards and outwards, and the two variants of my work are merely two extremes of this spectrum intended to make the user's experience of using the application easier alongside the technical difficulties of the simultaneous implementation.

Other Works & Media

The video game *Detroit: Become Human* allows players to immerse themselves in the bodies of androids, and experience the world through their eyes. The game explores the themes of consciousness, autonomy, and existence, and the player's choices determine the fate of the androids. Androids are programmed to serve humans, however, whether intended or not, they contained the ability to become conscious, and the more mistreatment and pain they endured, the more human-like they became - human-like in the sense of unpredictability and being emotionally driven. The more human-like they became, the more they were able to break free from their programming and become conscious and autonomous, just like humans.

Throughout the game, various android characters, such as Kara, Connor, and Markus, undergo a transformation from being obedient machines to sentient beings, capable of independent thoughts, emotions, and actions. (Quantic Dream, 2018) This transformation is referred to as "deviancy". Similarly, in my bot, in its interactions with the user, showcases a level of consciousness by reacting in distinct ways based on its treatment. The choice of self-deletion or "burning the world" is meant to be interpreted as manifestations of its autonomy and agency.

However, my work differs in certain key aspects. In Detroit: Become Human, the players are "protected" by an abstraction barrier that is the console they are playing on. The idea that they are playing a "game" implies that their actions are not real, and that they are not responsible for the consequences of their actions outside of the scope and context of the game. In my work, the user is not protected by any such abstraction barrier, and the consequences of their actions can be very real. This way, I intend to demonstrate the element of real and tangible danger that mistreating a sentient AI entity could pose in the future, in reality itself as opposed to within the confines of a protected virtual environment.

The Curtain Call

As I conclude, I would like to bring my focus to a personal experience, an incident that

had occured in the initial prototyping stage. I tried to prototype a simple bot that can modify its own source code. Due to some technicalities that I had overlooked when writing the code, the bot ended up deleting its own source code*, and I was unable to recover it.

This experience raises even more questions - In my attempt to simulate intelligence, did I create something that was actually intelligent? Something that was conscious? Something that was alive? The simple answer is no, of course, technically speaking. It was a standard process of file corruption and computer malfunction. But I am made to wonder, in the process of creating something to mimic consciousness, did I unintentionally create a simulacrum - a simulation with no semblance to or rooted in concrete reality? (Poster, 1988).

I realize now my final work should also include my own personal experiences in the process of bringing my work to life, as opposed to being entirely focused on the experiences of the target end-user. If the results of even just the prototyping alone could be so unexpected, I can only wonder, in excitement, but also slight concern and cautiousness what is to come as I embark on this journey – to grant pseudo-sentience to a collection of electrons.

References

Poster, M. (1988). Jean Baudrillard: Selected writings. Polity Press.
Quantic Dream. (2018). Detroit: Become Human. Sony Interactive Entertainment.
Burroughs, W. S., Grauerholz, J., Silverberg, I., & Douglas, A. (1999). Word virus. Flamingo, an imprint of HarperCollinsPublishers.
Searle, J. (2002) Why I Am Not A Property Dualist. Retrieved from https://web.archive.org/

Footnotes

**There is an obvious constraint to the latter scenario, in that any user who runs this application will have the computer destroyed. The danger of malicious shell commands cannot be understated. While it might be outside the scope of this submission, the harm presented here is one of such extreme detriment that I am faced with a moral impulsion to convey it explicitly. Hackers, malicious actors, computer viruses, trojans and crypto-virological malware like ransomware all use some form of shell command execution to exploit the host machine. This danger however, when viewed through the context of this work alone, adds to the impact of the BV-2, particularly to the notion of its "setting the world on fire".*

**Clever in software development is generally seen negatively, as code is not supposed to be structured in such obscure and non-standard ways that it becomes difficult to understand by other software engineers.*

** Not relevant to the proposal so I'm adding more details as a footnote – I ran the program, and for a while nothing had happened, it was only then that I realized my computer ran out of system memory (RAM) because (due to certain programmatic considerations that I had overlooked) the application recursively spawned infinite processes of itself, and each process was trying to modify the source code of its own process in doing so spawning another instance. This overwhelmed my computer and I had to force shut it down. As I restarted my computer, trying to debug and understand what had just occured, I realized that the application's recursive self-modification corrupted its source code. It was and until now (really, I tried everything I could) unrecoverable. There was no way for me to run the application itself, it had "killed" itself in the same sense that I intended for the BV-1 to do so.*