Of Ants and Men:

Essay on Mary Shelley's Frankenstein, or the Modern Prometheus

Jordy Pellemans

April 23, 2023

"Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

Max Tegmark, Life 3.0

"Trouble with mice is you always kill 'em."

Lennie Small, taken from John Steinbeck's 'Of Mice and Men'

Introduction

Imagine receiving a message from outer space that said, "People of Earth. We will arrive on your planet in fifty years. Get ready." It is a commonly heard imaginative inquiry in which the introduction of superintelligent AGI (artificial general intelligence) is rendered analogous to an extraterrestrial species casually informing us that they are 'en route' for Earth. Quite

obviously the metaphor was intentionally conceived of in a Hollywoodesque fashion to help us mere mortals have *any* idea of the sense of the urgency and unprecedented levels of uncertainty that may come about with an artificial intelligence explosion. And not so without reason, as nearly ten years ago, even one of the most remarkable scientific minds of our time, the late Stephen Hawking, critically remarked, "The rise of powerful AI will be either the best or the worst thing ever to happen to humanity. We do not yet know which (Hawking, 2014)."

In this essay I would like to expound on how AI-related themes feature richly in the 19th century novel *Frankenstein, or the Modern Prometheus* (henceforth: the Frankenstein novel) by Mary Shelley. I will then contend how this piece of exquisite proto sci-fi literature – even two centuries after it was first published – provides valuable insights, wisdoms and how it can enlighten us morally, ethically and philosophically with regard to the present-day developments in AI research and applications. In all, I hope to shed to some light on the question of 'How to deal with superintelligent AGI'?

Frankenstein, or the Modern Prometheus

The gothic novel of *Frankenstein* was written by Mary Wollstonecraft Shelley and published as early as 1818, at the beginning of the Industrial Revolution in Britain. Shelley, née Mary Wollstonecraft Godwin, was barely nineteen years old at the time. Although various accounts are in existence, it is widely believed that she came up with the idea for the novel while staying in a villa by Lake Geneva together with her husband, Percy Shelley and Lord Byron, and that due to the inclement weather conditions, they were forced to stay indoors and narrate ghost stories by the fireplace (Johnson, 1982; Shelley M. , 2008). In Shelley's journal entries, there is

one single passage in which the philosophical foundations and contours for her novel are clearly manifest:

> "Many and long were the conversations between Lord Byron and Shelley [i.e. Mary's husband], to which I was a devout but nearly silent listener. During one of these, various philosophical doctrines were discussed, and among others the nature of the principle of life, and whether there was any probability of its ever being discovered and communicated. They talked of the experiments of Dr. Darwin (I speak not of what the doctor really did, or said that he did, but, as more to my purpose, of what was then spoken of as having been done by him), who preserved a piece of vermicelli in a glass case, till by some extraordinary means it began to move with voluntary motion. Not thus, after all, would life be given. Perhaps a corpse would be re-animated; galvanism had given token of such things: perhaps the component parts of a creature might be manufactured, brought together, and endued with vital warmth."
>
> (Shelley M. W., 1987, p. 34)

Some scholars believe that, due to her social background, the socio-political rebellious nature of her parents and the backdrop of the Industrial Revolution gaining momentum, a lot of Shelley's writings were influenced by the 19th century Luddite movement (Mellor A. , 1988; O'Flinn,

1989; Butler, 2010), a group of workers in the textile industry who vehemently revolted against the rise of steam power-driven production at the very expense of artisanry (Hobsbawm, 1952).

The story of Frankenstein sets off with a series of letters from Robert Walton to his sister Margaret Saville. Walton is a sailor and describes his voyage to the North Pole, only to encounter a man by the name of Victor Frankenstein who is pitifully stranded there. Victor tells Walton his life's story including his fixation on reanimating dead beings. One of the beings he creates is indeed successfully revived, but he had not expected to feel so horrified by the creature's appearance. Consequently, Victor starts to severely doubt the morality of his act of bringing dead matter to life. *"But now that I had finished, the beauty of the dream vanished, and breathless horror and disgust filled my heart* (Shelley M. , 2008)." The creature is then mercilessly left to its own devices and comes to resent its maker for this very reason. It seeks revenge by killing those closest to Victor. The story ends with Victor being chased by the monster all the way up to the Artic, where both of them perish.

On Artificial Intelligence

Artificial intelligence, or AI, is the creation of computer systems that are able to mimic human-like intelligence, such as understanding natural languages, learning things and solving problems (Russell S. J., 2010; Goodfellow, 2016). For the purpose of this essay, it is first and foremost, crucial to distinguish between weak and strong AI. The former, also known as narrow AI, is capable of performing specific tasks which are normally solely the cognitive domain of humans. The latter, the theorized *strong* AI, also commonly referred to as AGI (artificial general intelligence), however, is, at this time of writing, not yet in existence, but its essence is markedly different in the sense that it is hypothesized to be able to perform virtually any task that humans

are capable of and even beyond. The significance of the word ‘beyond’ should, however, not be underestimated. For one, because predictions are that, once AGI is ‘given birth to’, and provided that it will be ‘allowed’ to autonomously keep reiterating, redesigning itself and accumulate more and more data and technological resources, we will fairly quickly – conceivably within our lifetimes – be passing into an age of massive uncertainty. That is, *uncertainty* in the sense that possibly even our brightest (human) minds would pretty much no longer be able to make even a single prediction about what could happen next. At this point, the metaphor of a black hole’s event horizon ‘singularity’ comes to mind, which is ‘a boundary around a black hole beyond which nothing, including light, can escape due to the intense gravitational pull (Narayan & McClintock, 2013).’ The idea of a technological singularity is a hotly debated theoretical concept that gained extensive popularity, respectively, through the essays and books by the pen of researchers Vernor Vinge and Ray Kurzweil (Vinge, 1993; Kurzweil, 2005). However, ultimately their ideas can be traced back to British mathematician Irving John Good, who first proposed the more foundational notion of ultra-intelligent machines (Good, 1965).

> “Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’ and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.”
>
> (Good, 1965, p. 67)

Ultimately, the bottom line here is that, once an intelligence explosion of AGI has been 'allowed' to set off, machine intelligence would increase exponentially. In the wake of this event, mathematically speaking, i.e. given the nature of exponentiation, the AI's intelligence level would accelerate and increase indefinitely. And this would pretty much lead to human beings no longer being the brainiest entity on the planet. Computer scientist and philosopher Stuart Russel puts it perfectly: "And just as ants cannot comprehend human accomplishments, we may be similarly [in the not-too-distant future] limited in our understanding of AIs (Russel, 2019, p. 2)."

The Light and the Fire

Let us now turn to the theme of fire that is featured in the Frankenstein novel and how it is related to AI. Firstly, the novel's secondary title inconspicuously hints at a connection between the ancient Greek myth of Prometheus:

> "For I did not think it right that mortals should have the power of the flame unaccompanied by all the other arts which belong to Hephaestus. And, in one word, I summed it all up, and, granting these things, I caused mortals to have a sense of shared community."
>
> (Aeschylus, 2008, pp. lines 447-452)

As it was Prometheus' plight to bring the fire and the light to humans, so it was Victor's goal to instill life into lifeless matter. Similarly, developments in AI research with regard to the creation of conscious AI which may or may not possess sentient agency have many times been thoroughly debated. Many controversies still surround the philosophical question of whether non-biological matter, complexly rearranged and restructured, may eventually be called self-

conscious and self-aware (Chalmers, 1995; Bostrom, 2014; Coeckelbergh, 2018). This idea of bringing inanimate things to life is anything but new and people's ambitions to achieve suchlike novelty have, so it seems, always been one of mankind's boldest dreams: to not be the created but to, instead, be the creator.

The theme of the light and the fire also reflects the centrality of scientific discovery. This is perhaps obvious from the aforementioned quote by Mary Shelley where she refers to experiments carried out by Dr. Darwin. Drawing on this metaphor, the fire represents the celebrated scientific method and its potential to enlighten our minds. Extending the metaphor of SCIENCE is FIRE, we may also reason that fire offers us light, warmth and protection. Whenever the fire goes out, it will leave us in darkness. But oppositely, too intense a fire is hazardous, and similarly, too much light may be dangerously blinding. In this manner the wandering Monster experienced fire in diametrically opposing ways: 'One day, when I was oppressed by cold, I found a fire which had been left by some wandering beggars and was overcome with delight at the warmth I experienced from it. In my joy I thrust my hand into the live embers, but quickly drew it out again with a cry of pain. How strange, I thought, that the same cause should produce such opposite effects!' (Shelley M. , 2008, p. 148) Using trial-and-error, the Monster interacts with the outside world, and it is striking how *fire*, providing warmth and brightness, has such painful consequences when touched, despite the appealing outside characteristics. In her essay 'The spark of life: Science and the electrical metaphor in Mary Shelley's Frankenstein', Mellor, too, recognizes and acknowledges Shelley's cautionary stance. "By identifying Victor's experiment with the destruction of the natural order and the elevation of human ambition over natural limitations, Shelley critiques the modern scientific enterprise as a Promethean act of arrogance that can only lead to the destruction of the human and natural

worlds (Mellor A. , 1988, p. 235)." In addition, the reference to the ‘Ancient Mariner’ (Coleridge, 1834) in that respect offers a beautiful instance of intertextuality and introduces analogies that foreshadow the story’s dramatic outcome (Shelley M. , 2008, p. 12). Relating this back to AI, this inevitably begs the question of: are we human beings, in our endeavors and zeal to go beyond boundaries, being arrogant? If we can, does this mean that we should? One could argue this to be arrogant and selfish, although it may also arguably be a natural stage of evolution. In his book, ‘Life 3.0’, Swedish-American physicist Max Tegmark proposes ‘The Three Stages of life’. Firstly, Life 1.0 means being capable of surviving and replicating itself (simple biological). Life 2.0 is capable of designing its own software (cultural). And, finally, Life 3.0 is capable of designing its own hardware (technological). The point here is that, according to Tegmark, with the rise of AI, our species is on the verge of a technological era. ‘Life 3.0 is the master of its own destiny, finally fully free from its evolutionary shackles (Tegmark, 2017).’ While his book is on the tentatively optimistic, non-Luddite side of things with respect to AI, he does clearly recognize the potential dangers of human overconfidence in our ability to create and control advanced AI. Conversely, in the Frankenstein novel’s beginning, solely scientific optimism seems to reign supreme. For example, the novel features a very telling passage where Victor is lectured by one of his professors, M. Waldman.

> “The ancient teachers of this science,” said he, “promised impossibilities and performed nothing. The modern masters promise very little; they know that metals cannot be transmuted and that the elixir of life is a chimera. But these philosophers, whose hands seem only made to dabble in dirt, and their eyes to

> pore over the microscope or crucible, have indeed performed miracles. They penetrate into the recesses of nature and show how she works in her hiding-places. They ascend into the heavens; they have discovered how the blood circulates, and the nature of the air we breathe. They have acquired new and almost unlimited powers; they can command the thunders of heaven, mimic the earthquake, and even mock the invisible world with its own shadows."
>
> (Shelley M. , 2008, p. 53)

The professor makes the case for 19th century scientists and their modest but empirical methodology, as opposed to the alchemists' methods which were largely based on rather superstitious, mystical and proto-scientific foundations broadly incompatible with modern science (Newman, 2004). Victor is so enthralled by the professor's speech that he is instilled with a deep sense of purpose:

> I felt as if my soul were grappling with a palpable enemy; one by one the various keys were touched which formed the mechanism of my being; chord after chord was sounded, and soon my mind was filled with one thought, one conception, one purpose.
>
> (Shelley M. , 2008, p. 54)

This passage perhaps underlines the idea that a craving for the unknown and exploring the unexplored can lead people to be blinded. Our unquenchable thirst for knowledge and 'knowing' is like the proverbial 'moth to a flame'. We are completely mesmerized, having only one goal in mind and yet, *hoc momento*, we fail completely to reason effectively about the intended and unintended outcomes. Interestingly, Victor, after a restless night, then visits M. Waldman privately and the two discuss Victor's ideas. It is then the professor who says:

> The labours of men of genius, however erroneously directed,
>
> scarcely ever fail in ultimately turning to the solid advantage of mankind.
>
> (Shelley M. , 2008)

It the above passage, Shelley, by means of the two men's larger-than-life overconfidence and Promethean philosophy, antithetically foreshadows the gloomy outcomes of Victor's experiments and his ultimate fate.

> "Thus ended a day memorable to me; it decided my future destiny (Shelley M. , 2008)."

Alignment, Control and Parental Responsibility

We have so far looked at scientific and ethical responsibility. With regard to AI, however, Shelley's novel seems to have more up its literary sleeve. These are on the one hand what is called the 'alignment problem' and 'parental responsibility'. I will explain what they are

and how they are inextricably linked together.

Firstly, the ‘alignment problem’ is a term broadly known among AI researchers and experts to refer to the challenge of making AI do what we want without getting ourselves wiped out in the process. More scientifically phrased, “the AI alignment problem is about ensuring that an AI's behavior is aligned with what we really want. It requires ensuring that the AI behaves safely and does not cause unintended consequences. It also requires ensuring that the AI is robust to changes in its environment, including changes in the goals it is intended to pursue (Orseau & Armstrong, 2016, p. 1)." In some sense it is deeply ironic that the greatest threat in creating superintelligent AI, including robotics – the branch of AI which deals with mastering locomotion – is not that AI does not do what we *want* it to do. Instead, the real danger, as real as can be, is that AI will do *exactly* what we want it do. That is, that it will do so without implicitly keeping in mind human values and world views. For example, if in the foreseeable future a superintelligent AGI were asked to solve climate change, it may very well decide that the most effective way to do so would be to rid our planet of us, i.e., human beings, entirely. Actual examples of AI not behaving as expected are actually well documented and numerous books and articles were written on the problem (Bostrom, Superintelligence: Paths, Dangers, Strategies, 2014; Russel, 2019; Yampolskiy, 2019). With the incredibly recent introduction of so-called automated generative pre-trained transformers (autoGPTs) – which possess a form of interactive agency – mankind once more finds itself one significant step closer to inventing true AGI.

Turning to the Frankenstein novel, the theme of ‘control’ is ubiquitous throughout the novel. The Promethean philosophy adopted by Victor which ultimately leads him to nearly solipsistically create life is ultimately about control. For us humans, it was first through the evolution of our brain’s neocortex that we gained the ability to work together more effectively

and outsmart and thus ‘control’ our environment. Then, it was through our cultural development and the passing on of rituals, wisdoms and information that we gained even more ‘control’. And now, in our present-day reality, when considering advances in AI, we again find ourselves on the edge of potentially – hopefully - acquiring even more ‘control’. However, the extent of this control is fully dependent on whether or not we will be able to harness AI alignment. Revisiting the novel again, Judith Butler keenly observes that Victor realizes that, by instilling life to dead matter, and succeeding, he may have become the monster himself, and with the monster running off, he has lost control.

> “Something ran off, got out of control, at the moment when the title gave birth to a character by that same name who left his home only to create a monster who roams widely through the landscape and seems to share the name of its creator. Indeed, the title is animated by both the scientist and his monstrous “progeny,” naming a certain form of creation that loses control of what is created.”
>
> (Johnson, 1982, pp. 39-40)

In this respect, the act of creation may very well tantamount to losing control. For example, when as human beings we procreate and become parents, we by definition lose some control, as the goal of having children is inevitably to one day let them go and have them live their lives both independently and autonomously. Victor, in creating his monster, has foregone thinking about what it would be like to create a new soul, including their personality, needs and desires. Julia Cameron puts this nicely when saying that ‘The creative process is a process of surrender, not control (Cameron, 1992).” Although originally intended to characterize a creative art process, Cameron’s idea is very much applicable to Victor’s attitude toward the monster as well. Victor was never ready to be a ‘parent’ let alone having given even a single thought about

the needs of his creation. At one point, however, Victor does indeed seem to indulge in creating a female partner for the creature, only to destroy her before bringing her to life:

> The wretch saw me destroy the creature on whose future existence he depended for happiness, and with a howl of devilish despair and revenge, withdrew.
>
> (Shelley M. , 2008, p. 252)

It seems that Victor had one goal in mind and one goal only: to create life. The moral and ethical responsibilities of the act are overlooked, or perhaps, they were never considered in the first place. He seemingly never had any intention of raising a child, but when it was there, he treated it horribly and, as a result, the whole thing severely backfired on him. In his book, Scary Smart, author and former supervisor of GoogleX, Mo Gawdat, tells the well-known story of a certain Cryptonian endowed with superhuman powers, who lands on Earth as an infant, grows up and later goes by the name of Superman. Gawdat then compares Clark Kent to AI as an infant. Because, if AI is still in its infancy, and the adoption paperwork has already been filled out, then we ought to make sure to do some excellent parenting and act as role models.

Conclusion

In conclusion, our world is expecting; that is expecting something we conceived of us as a race, together. Shelley's Frankenstein novel provides helpful insights and teaches us that the act of creation comes with inevitable ethical and moral responsibility. To unearth new knowledge about this world, is one thing. But to create novel things based on that knowledge, we must be aware that certain risks are involved, i.e., the outcome may not be as romantic as we had hoped it to be, and we might lose control over it. Similar to Victor's goal to revive lifeless

matter, developments in AI are currently on track to create superintelligent AI, spawning a race of AIs, which may or may not be attuned to our precious human values. Whether it will or will not, may very well depend on how we, as human beings, treat each other on a daily basis. As human beings, we may have to get ourselves ready for parenting AI.

Works Cited

Aeschylus. (2008). *Prometheus Bound.* (T. b. Smyth, Trans.) Harvard University Press.

Bloom, H. (2007). *Bloom's Modern Critical Interpretations: Frankenstein, Updated Version.* Infobase Publishing.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Bostrom, N. (2014). The idea that eats smart people. *Journal of Evolution and Technology*, 1-22.

Butler, M. J. (2010). *Literature, Frankenstein and Radical Science. In Body of Writing: Figuring Desire in Spanish American.* Duke University Press.

Cameron, J. (1992). *The artist's way: A spiritual path to higher creativity.* TarcherPerigee.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 200-219.

Cho, S. (2016, March 9). Google's AlphaGo beats Lee Sedol in Go, the complex game that is 'hard to encode. *The Washington Post*. Retrieved from Google's AlphaGo beats Lee Sedol in Go, the complex game that is 'hard to encode.:

https://www.washingtonpost.com/news/morning-mix/wp/2016/03/09/googles-alphago-beats-lee-sedol-in-go-the-complex-game-thats-hard-to-encode/

Coeckelbergh, M. (2018). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, 207-223.

Coleridge, S. T. (1834). *The Rime of the Ancient Mariner.*

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In F. Alt, & M. R. (Eds.), *Advances in Computers, Vol. 6* (pp. 31-88). Academic Press.

Goodfellow, I. B. (2016). *Deep learning.* MIT Press.

Hawking, S. (2014, December 2). Stephen Hawking warns artificial intelligence could end mankind. (B. News, Interviewer)

Hobsbawm, E. J. (1952). The machine breakers. *Past & Present*, 57-70.

Johnson, B. (1982). *A life with Mary Shelley.* Harvard University Press.

Jordan, M. I. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 255-260.

Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology.* New York: Viking.

Mellor, A. (1988). *Possessing Nature: The Female in Frankenstein.* Indiana University Press.

Mellor, A. K. (1988). The spark of life: Science and the electrical metaphor in Mary Shelley's Frankenstein. *Studies in Romanticism*, 227-255.

Narayan, R., & McClintock, J. E. (2013). Observational evidence for black holes. In S. S. Holt, & C. D. (Eds.), *The Astronomy and Astrophysics Decadal Survey, volume 2010* (pp. 39-54). The National Academies Press.

Newman, W. R. (2004). The secrets of alchemy. *The Chemical Heritage Magazine*, pp. 26-31.

O'Flinn, P. (1989). *Production and Reproduction: The Case of Frankenstein.* Palgrave Macmillan UK.

Orseau, L., & Armstrong, S. (2016). Safely Interruptible Agent. *Journal of Artificial Intelligence Research*, 1-99.

Ortiz, S. (2023, April 14). *What is Auto-GPT? Everything to know about the next powerful AI tool*. Retrieved from www.zdenet.com: https://www.zdnet.com/article/what-is-auto-gpt-everything-to-know-about-the-next-powerful-ai-tool/

Russel, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Penguin Random House.

Russell, S. J. (2010). *Artificial Intelligence: A modern approach.* Prentice Hall.

Shelley, M. (2008). *Frankenstein.* Oxford University Press.

Shelley, M. W. (1987). *The Journals of Mary Shelley, 1844-1844* (Vol. II). (P. R. Feldman, & D. Scott-Kilvert, Eds.) Oxford, England, England: Oxford University Press.

Shelley, P. B. (2021). *Prometheus Unbound - A Variorum Edition [Adobe Digital Editions version.* University of Oklahoma Press.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.* New York, USA: Alfred A. Knopf.

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. *Whole Earth Review*(81), pp. 22-27.

Yampolskiy, R. (2019). *AI Alignment: Why It Matters and How We Might Achieve It.* Oxford University Press.