Working Paper No. 1073

Frankenstein in Fact and Fiction

Lecture at Bard College, November 19, 2024

by

Lord Robert Skidelsky
Warwick University

December 2024

Levy Economics Institute
P.O. Box 5000
Annandale-on-Hudson, NY 12504-5000
http://www.levyinstitute.org

ISSN 1547-366X

INTRODUCTION

As we all know, Frankenstein was the scientist in Mary Shelley's 1818 novel of the same name, who invented a human machine—intended to be a benefactor, but which turned out to be a monster. There is a critical question I wish to address this evening: Can we avoid our technology destroying us? This is the most important thread that runs through my book, *Mindless* (2024), recently published in the United States. The book discusses the impact of machines on jobs, on freedom, and on our survival as a species. The question that dominates all three concerns the impact of machines on our humanness. Today we ponder whether there is still time to control the Machine before it controls us. I will talk about three Frankensteins who each set out to create gods and, in turn, created monsters. Let me start with Mary Shelley.

I. MARY SHELLEY'S *FRANKENSTEIN*

Mary Shelley's *Frankenstein; or, The Modern Prometheus* (1818) marks the historical moment when the Enlightenment dream turned sour. It tells the story of a hominoid creature, constructed from dead body parts and animated by a battery, to serve its idealistic master—the medical scientist Victor Frankenstein, who wants to create a more perfect human. His creation, though, turned on its creator and became a killing machine. Shelley's story of a science experiment gone wrong became a metaphor for the whole endeavor of mechanizing human intelligence, and marks the start of the humanistic revolt against the machine, a warning against hubris, which, alas, has not been heeded.

Mary was married to the poet Percy Bysshe Shelley, whose dramatic poem "Prometheus Unbound" (1820), with its message of the liberating power of science, was penned at exactly the same time as Mary wrote Frankenstein. Percy's Prometheus is the ideal revolutionary: he rebels in the name of humanity, impelled, the author says, "by the purest and truest motives to the best and noblest ends." Mary Shelley tells a very different story. Her *Modern Prometheus* reflects the growing opposition of Romantic poets to the horrors of the Industrial Revolution.

One can only wonder at the marital disharmony between the Shelleys which must have resulted from two such contrasting views of humanity's future. Let me also offer this thought: Prometheus is a peculiarly western myth—born in classical Greece, taken over by the

Enlightenment. We don't find the story's equivalent in other civilizations. This should influence the way we think about the future of science and technology.

The novel's protagonist, Victor Frankenstein, is a medical doctor, skilled in the surgery of dissection, who seeks to learn how to create life from inanimate matter. In conceiving him, Mary Shelley drew both on the traditional literature of animated puppets and automated robots, and on the new materialist science in which consciousness is a byproduct of matter. The science of her *Modern Prometheus* reflects the debates on galvanism (the use of electricity to stimulate or restart life) taking place at the time.

We must keep this aspect—the dream of creating conscious life from inanimate matter—in mind, as it continues to inspire the effort to create a modern Prometheus. But Frankenstein's creature never attains the godhood promised by Enlightenment science, turning instead into a monster. The novel thus has the double character of a Gothic horror story and of science gone horribly wrong.

Early on, Victor Frankenstein is motivated in his scientific endeavors by a mixture of ambition and humanitarianism: '"what glory would attend the discovery if I could banish disease from the human frame and render man invulnerable to any but a violent death!'" All Frankensteins combine a genuine benevolence with the quest for power and glory. As the novel progresses, the quest for power takes control. Frankenstein becomes possessed by a demonic hunger for power over nature: "'with unrelaxed and breathless eagerness, I pursued nature to her hiding-places […] A new species would bless me as its creator and source; many happy and excellent natures would owe their being to me. No father could claim the gratitude of his child so completely as I should deserve theirs.'"

There has been much debate about why Frankenstein's creature turned into a monster. One theory, in line with modern psychology, is that the creature was hated by its creator because it was repulsive to behold. The unloved child became the criminal. If only Viktor had cared for it—by, for example, creating a wife for it—all would have been well.

At the end of the novel, a dying Victor Frankenstein entreats the narrator who tells his tale, to "seek happiness in tranquillity and avoid ambition, even if it be only the apparently innocent one of distinguishing yourself in science and discoveries.” But, of course, “science and discoveries” can never be innocent.

In Mary Shelley’s *Frankenstein*, one can find recognizable features of the scientific/ technological spirit of our own time, as it strives to create an inhuman humanity.

II. A MODERN FRANKENSTEIN: JOHN VON NEUMANN

The first half of the twentieth century produced several real-life Frankensteins. The most remarkable, in his genius, ambitions, and effects was the Hungarian mathematician John von Neumann (1903–57).

A book about Neumann, titled *The MANIAC* (2023) has recently been published, written by South American Benjamin Lapatut. I wish I had had it before I wrote *Mindless*.

Johnny Neumann was a mathematical prodigy. He invented the mathematics of quantum mechanics, of the electronic computer, and of the atomic bomb. With Oskar Morgenstern he invented game theory, with applications in both economics and military strategy. He foresaw the so-called singularity—the moment at which machines surpass humans in everything humans do—and he promised godlike control over the Earth’s climate.

What fascinates me about the Neumann story is its mixture of divine and diabolic: his genuine belief in mathematics as the path to salvation coupled with his acceptance of evil as the agent of good, and his view of technology, however destructive, as the price of progress, which led him to turn his maths toward creating weapons of mass destruction. His life sums up the spirit of the machine age, which saw in science and technology the only antidote to the human wickedness revealed in the two world wars. It barely occurred to them that they might be fathering even more extreme forms of madness.

Putting it very simply, an influential group of mathematicians and scientists, largely Hungarian—why Hungary I ask myself?—saw in mathematics the only answer to human irrationality. Maths would put reason in uncontestable control of human affairs. Since

mathematics promised a method of proving or falsifying any proposition, there would be no space for the windy militant trash which drove populations to orgies of mass killing.
The quest for the certainty which could both legitimize some behaviors and eliminate others had a profound effect on twentieth century philosophy, politics, ethics, and economics. Karl Popper's *The Logic of Scientific Discovery* (1934) is perhaps the best-known application of the falsification principle to both the natural and human sciences. The quest to eliminate partial narratives and "fake news" from public discourse is a notable feature of our own day, even as social media provide ever more opportunities to express them.

More broadly, our modern Frankensteins emerged from an intellectual atmosphere in which old religion had broken down, but the need for religion had not. Lost faith left a gaping hole which needed to be filled. Communism—Arthur Koestler's *God That Failed* (Spender et al. 1949)—is the most famous example of a creed which viewed itself simply as social mathematics. Zamayatin's great dystopian novel *W*e (1920) pictures a political system, the legitimacy of which rests on the truths of maths.

Following the lead of the German mathematician David Hilbert, Neumann argued that a mathematical program to control society could be built on a finite set of axioms and rules of inference, as solid as those of geometry. This foundation would make possible a non-contradictory social logic: a secure basis for all thought and action.

At Konigsberg in Sept 1930, the Hilbert-Neumann dream of salvation through maths was torpedoed by the graduate student Kurt Godel, who stammered out what came to be called his incompleteness theorem before the assembled mathematical luminaries. Godel showed that no formal system can be both complete and consistent. One could not prove the truth of maths by maths. Some bits of maths were undoubtedly true, but unprovable; so consistency—freedom from contradiction—cannot be guaranteed by mathematics.

Following Godel's demolition of his ideal project, Neumann "became a renegade mathematician, a mind for hire, increasingly seduced by power" (Labatut 2023). Emigrating to the United States, it became his overriding purpose to apply mathematics to weapons. The aim of course was benevolent: to save the world from Hitler, subsequently from Stalin. The method, though, now explicitly involved the acceptance of evil as a means to good. Neumann was one of the mathematicians shipped off to a secret laboratory in New Mexico which housed the Manhattan Project.

After the war, Neumann continued working on what Einstein called “the great technologies of death.” This led to famous academic rows with Einstein and Oppenheimer. Einstein wanted to stop further development of atomic weapons. Neumann agreed that the scientists were creating a monster. But they had to go on with it, not just for military but for ethical reasons. “It would be unethical,” Neumann said, for scientists “not to advance what they know is feasible, no matter what terrible consequences it may have” (quoted in Labatut [2023]).

Like that other lapsed mathematician Bertrand Russell, Neumann advocated a surprise attack on the Soviet Union. Permanent peace—an undoubted good—required we rain nuclear hell on the Russians before they developed weapons of their own. A Pax Americana would be built on a heap of corpses. A nuclear first strike before the enemy had nuclear weapons was the only rational course of action.

In his book, *Theory of Games and Economic Behaviour* (1944) co-authored with Oskar Morgenstern, Neumann provided the maths for the deadly game of Mutually Assured Destruction (MAD), a game of chicken played out on a planetary scale with weapons capable of global destruction.

The flaw in game theory is obvious enough: it presupposes perfectly rational and logical agents, interested only in winning (maximizing their utilities), who possess a perfect understanding of the rules, a total recall of all past moves, and a flawless awareness of the ramifications of their own and of their opponents’ actions at every single stop in the game. It did not take a Keynes to point out that this is not normal. He based his economics on a denial of all these propositions, but we still teach economics as though it were a “game” in the Neumann sense.

In 1946, Neumann promised the US military he would build them a computer powerful enough to handle the intricate calculations needed for the creation of a hydrogen bomb. The machine was the Mathematical Analyzer, Numerical Integrator and Computer: MANIAC. Its first use was to provide a mathematical program for thermonuclear weapons.

Looking at the history of our modern Frankensteins, it is remarkable how both the most creative and the most destructive of human inventions arose at exactly the same time in history. As Richard Feynman (quoted in Labatut [2023]) has pointed out, “[s]o much of the high-tech world we live in today, with its conquest of space and extraordinary advances in biology and medicine, were spurred on by one man's monomania and the need to develop electronic computers to calculate whether an H bomb could be built or not.”

Yet, had he lived into our own times, Neumann would have been able to justify his work by saying that for seventy years or so, between them, MAD and MANIAC held in check the greater madness of a thermonuclear war.

The second goal Neumann set for MANIAC was to develop new kinds of life. He said: “You insist that there is something a machine cannot do. If you tell me precisely what it is a machine cannot do, then I can always make a machine which will do just that.” The workshop held at Dartmouth College in New Hampshire in the summer of 1956 is seen as the birth of artificial intelligence (AI). Its proposition that “every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it” was pure Neumann.

In 1955, Neumann was appointed by Eisenhower as one of six heads of the US Atomic Energy Commission. In his first paper, Neumann proposed using hydrogen bombs to divert the path of hurricanes: “All processes that are stable we shall predict, all processes that are unstable we shall control.”

In February 1957, Neumann died in the Walter Reed Army Medical Centre, surrounded by military types desperate to catch his last words on how to get the better of the Soviets. In one of his last letters, he wrote: “Cavemen created the gods. I see no reason why we shouldn’t do the same.” The modern world needed new gods to bridge the divide between thought and computation. He wrote, “if we could understand the primordial language of the brain, it would transform the prospects of mankind. Perhaps some day we could begin to merge with machines.” He never thought of blocking the road to merger. “For progress there is no cure,” he said.

III. TRANSHUMANISM OR FRANKENSTEIN TODAY

Today's Frankensteins are a mixture of high-tech multi-billionaires, scientists, and philosophers of science united by the common goal of developing super-intelligent beings to save humanity from itself. They pursue "generative AI"—"a type of AI," Google tells us blandly "that can create new content…"; more understandably, this is a type of AI that can think for itself. Since its thoughts will not be burdened by human fallibility, it will be able to think *better* than humans, and thus behave more rationally. It is the latest iteration in the dream of creating God. Can we, however, be sure that the AI that thinks for itself will be a benevolent God? Some of the funders and scientists have doubts. In 2023, Elon Musk called for a six-month pause for reflection; others have asked for a longer pause before allowing the ascent to super intelligence. However, these are only pauses to allow us to develop fail-safe mechanisms. The God Project itself remains on track, seemingly irreversible.

It has acquired a new philosophical legitimacy called Transhumanism, which the philosopher Émile Torres (2021) sees as "quite possibly the most dangerous secular belief system in the world today."

Transhumanism's leading lights are three philosophers, William MacAskill, Nick Bostrom, and Toby Ord. Nick Bostrom is head of the Future of Humanity Institute at Oxford, which also houses Toby Ord's Centre for Effective Altruism. William Macaskill heads a Global Priorities Institute. Similar institutes and think tanks are to be found in other universities. Their creed is long-termism, their *modus operandi*, effective altruism. They preach escape from human to superhuman intelligence as the only way of salvaging what they think of as humanity. Their institutes are financed by high tech billionaires like Elon Musk and Peter Thiel.

Transhumanism's value is that it alerts us to the destructive power of modern technology, but is then led by a perverse logic to advocate a transhuman (super-intelligent) form of technology which is profoundly *inhuman*. Humanity is cast as a reified construction, identified with the property "intelligence," divorced from the actual life of people. This mistake of the Enlightenment is carried by transhumanism to the point of madness, providing our contemporary Frankensteins with new intellectual toys.

The philosophic starting point is utilitarianism. The goal of policy should be to maximize the utility of the universe. The rightness of an action is to be judged by its long-run consequences for utility. The end justifies the means, and no means are ruled out of court *ab initio.* This is the way the game is set up.

The next step on the road to madness follows from the logic of counting heads. It is the quantity of utility which matters, not quality. This means treating everyone's utility the same, including that of those yet unborn. Thus, the goal is not to maximize the utility of the present generation, but of all feasible future generations, of which this generation will form only a tiny fraction. Ethically speaking, the utility of our generation should make only a tiny claim on our moral concern. As Ord (2020) puts it: "because, in expectation, almost all of humanity's life lies in the future almost everything of value lies in the future as well." Effective (or impartial) altruism prioritizes the interests of the yet unborn over those of the present generation.

The next step in the argument identifies the goal of maximizing the utility of the universe with that of maximizing its intelligence *potential*—that is, its capacity for creating value. Humans are unique among animals in their cognitive ability. Their cognitive potential has steadily advanced through the operation of the Darwinian survival of the fittest, and with billions of survivors now inhabiting the planet, humanity's intelligence is exploding. Moore's Law about the exponential increase in computing power is an example. If intelligence is identified with big data and computation then there would seem to be no limit to machine learning.

As AIs grow in intelligence, they will take charge of evolution. Already AIs are being built which can match the best of human intelligence. It is more than likely, and sooner rather than later, that humans will be able to design superintelligent AIs. Bostrom (2014) defines superintelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest." Since the design of machines is one of these cognitive performances, modestly super-intelligent machines will design even better machines to replace them; there would then be "an intelligence explosion." A population of super-intelligent AIs would take over the business of evolution, leaving the intelligence of man far behind. Thus "the first ultra-intelligent machine is the last invention that man need ever make." The evolutionary torch will have passed from humans to AIs.

The logic grinds on remorselessly. The body depends on the finite resources of our planet. But super-brains would be able to detach themselves from the limitations of the body. They might then escape from the limitations of our world, and establish colonies in our planetary "light cone," to be "fed" from its still unexhausted "endowment of negentropy" (or reverse entropy) in our cosmos. Humanity's intelligence potential could then be preserved and expanded for millions of years until the sun has finally cooled. Actual humans are nothing but means to this end, and therefore valuable only insofar as they contribute to the overall net amount of value in the universe between the Big Bang and the heat death. This is the philosophic/moral basis of the billionaire-financed projects of escape to the moon and other planets.

At this point, eschatological urgency seizes control of the transhumanist argument. The transhumanists share the view of the doomsday scientists that AIs programmed with human intelligence only might quite possibly produce a nuclear or environmental catastrophe. Ord pays particular attention to "near misses" during the Cuban missile crisis of 1962. However, even such catastrophes need not be fatal to our intelligence potential, provided there are survivors. In this scenario, echoes of biblical prophecy are mixed with the more sinister eugenicist musings of Dr. Strangelove, the "mad" scientist of Stanley Kubrick's 1964 film of that name, who suggested preserving the best of humanity in deep mines following a nuclear catastrophe.

However, with superintelligent AIs in charge of even more potent weapons of mass destruction than those available today, there might well be no survivors, either human or artificial. Thus the coming of superintelligence offers the possibility of either immortality or total disaster. We aim to create a benevolent God, but it is always possible that, like in the Frankenstein story, this God may turn out to be a Deus Malignus, who might only *pretend* to have good intentions, but unchained, would set about destroying not only us but its AI rivals.

So before they finally take over, our superintelligent AIs must be programmed with moral rules. The only moral rules available, however, come from our own imperfect and conflicting moral values. Wriggle as they might, transhumanists cannot escape the dilemma that there is no possibility, in a world of value relativism, of binding super-intelligence to an agreed upon morality. To put it in transhumanist terms, why should super-intelligent AIs be effective altruists? To answer: So the benevolence of our future controllers cannot be guaranteed. While recognizing the risk to humanity of super-intelligent AIs, transhumanists are too entranced by their dream of a cosmic computronium to propose shutting AI down before it reaches super-intelligence. Thus, Ord (2021) writes: "a permanent freeze on technology…would probably itself be an existential catastrophe, preventing humanity from ever fulfilling its potential."

One can easily see how remorseless logic, unchecked by common sense and ordinary humanity, can lead to madness. All follows from the commitment to untrammelled utilitarianism. Not only does this rest on hubristic claims about our ability to predict the future effects of our actions (the transhumanists assume a "perfect Bayesian calculator"), but it is deeply corrupting, insofar as it tempts us to override the common decencies of life in the name of an abstract future good. William Blake put it well: "He who would do good to another must do it in Minute Particulars: general Good is the plea of the scoundrel, hypocrite, and flatterer." Utilitarianism's object of concern is not "my neighbour"—that is, the concrete other who confronts me—but the abstract, individual humanity. The idea of concentrating funds and research efforts to maximize the unactualized possibility of intelligence throughout eternity is an extreme (insane) form of a Satanic project.

CONCLUSION

At issue in the debate about automation is what it is to be human. Today's scientists and technologists follow in the footsteps of Victor Frankenstein and Johann von Neumann in aiming to develop machines which can improve upon humanity. The result is a peculiarly anti-human form of humanism.

Take, for instance, the debate about up-skilling. We need to be constantly upskilling humans to be able to race with the machines; only that way will we retain our place in a machine culture. But what sort of place will this be?

Summarizing the work of Shannon Vallor (2021), Nathan Gardels (2021) describes the issue as follows:

> While algorithms can outperform humans in manifold tasks, as well as learn new ones. they literally do not understand what they are doing. Understanding comes from context. The uniquely human labor of filling in the cracks between bits of data with unprogrammable awareness is what creates meaning and constitutes a whole reality. Yet the more our minds are trained by daily interactions with digital technologies to think like algorithms that lack understanding, the less intelligent and more artificial we ourselves become.

From its inception, the Frankenstein project has rested on a materialist interpretation of the mind. Victor Frankenstein formed his thinking creature from body parts. Descartes believed that the "soul" (or mind as we now call it) was located in the pineal gland. The hunt for the "ghost in the machine" has continued ever since. Scientific inquirers sliced up the dead Einstein's brain hoping to find in it the seat of his genius. The modern consensus is that consciousness exists, and is unique to humans, but is simply a neurochemical structure of exceptional complexity. In understanding how the brain works, one will have cracked the secret of consciousness. It then becomes only a matter of time before science can breed thinking brains outside the human body. At some point, the human and the Divine will meet.

But the older, and historically far more influential view, is that humanity is constituted by its imperfection. Evil is a necessary part of the "great chain of being," the physical and moral challenge God has set to humanity. If all low and evil conditions and characters are eliminated, the harmony of the whole is irretrievably spoiled.

Today we need to be concerned not about the possibility of conscious AIs, but about the threat posed by the AIs we already have or know how to make.

I leave you with these questions. Can we stop certain lines of scientific enquiry by not funding them? Is everything invented bound to be applied? Is it possible to confine scientific advances to cases where the benefits are so clear as to be indisputable, e.g., medical advances. As soon as one says this, however, one realizes that there are no unmitigated benefits or ways of avoiding arbitrarily balancing the good and the bad. How this balance is determined will depend on our view of what it is to be human, that is, ultimately on a religious view of life's meaning and purpose.

REFERENCES

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

Gardels, N. 2021. “AI Makes Us Less Intelligent and More Artificial.” *Noēma,* February 5, 2021. https://www.noemamag.com/ai-makes-us-less-intelligent-and-more-artificial/

Labatut, B. 2023. *The MANIAC.* New York: Penguin Press.

Neumann, J. and O. Morgenstern. 1944. *Theory of Games and Economic Behaviour.* Princeton: Princeton University Press.

Ord, T. 2020. *The Precipice: Existential Risk and the Future of Humanity.* New York: Hachette Books.

Popper, K. R. (1934). *The Logic of Scientific Discovery*. Eastford, CT: Martino Fine Books.

Shelley, M. 1998 (1818). *Frankenstein; or, The Modern Prometheus.* Oxford: Oxford University Press.

Shelley, P. B. 1820. “Prometheus Unbound.”

Skidelsky, R. 2023. *Mindless: The Human Condition in the Age of Artificial Intelligence.* New York: Other Press.

Spender, S., L. Fischer, A. Koestler, I. Silone, R. Wright, and A. Gide. 1949. *The God that Failed.* New York: Harper and Row Publishers.

Torres, E. P. 2021. “Against Longtermism.” *Aeon,* October 19, 2021. https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo

Vallor, S. 2021. “The Thoughts the Civilized Keep.” *Noēma,* February 2, 2021. https://www.noemamag.com/the-thoughts-the-civilized-keep/

Zamyatin, Y. 1924. *We.* New York: Penguin Random House.