FILMS FROM THE FUTURE

The Technology and Morality of Sci-Fi Movies

Andrew D. Maynard

Mango Publishing
Coral Gables

Cover & Layout Design: Jermaine Lau

For permission requests, please contact the publisher at:

Mango Publishing Group
2850 Douglas Road, 3rd Floor
Coral Gables, FL 33134 USA
info@mango.bz

For special orders, quantity sales, course adoptions and corporate sales, please email the publisher at sales@mango.bz. For trade and wholesale sales, please contact Ingram Publisher Services at customer.service@ingramcontent.com or +1.800.509.4887.

Films from the Future: The Technology and Morality of Sci-Fi Movies
Library of Congress Cataloging
ISBN: (print) 978-1-63353-907-5 (ebook) 978-1-63353-906-8

Library of Congress Control Number: 2018955574

BISAC category code: SCI075000—SCIENCE / Philosophy & Social Aspects

Printed in the United States of America

CHAPTER EIGHT

EX MACHINA: AI AND THE ART OF MANIPULATION

> *"One day the AIs are going to look back on us the same way we look at fossil skeletons on the plains of Africa. An upright ape living in dust with crude language and tools, all set for extinction."*
>
> —Nathan Bateman

Plato's Cave

Over two millennia ago, the Greek philosopher Plato wrote *The Republic*. It's a book that continues to be widely influential. And while it's not widely known for its insights into advanced technologies, it's a book that, nevertheless, resonates deeply through the movie *Ex Machina*.

Like *Ghost in the Shell* (chapter seven), *Ex Machina* explores the future emergence of fully autonomous AI. But unlike *Ghost*, the movie develops a plausible narrative that is set in the near future. And it offers a glimpse that is simultaneously thrilling and frightening into what a future fully autonomous AI might look like. Forget the dystopian worlds of super-intelligent AIs depicted in movies like *The Terminator*,[101] *Ex Machina* is far more chilling because it exposes how what makes us human could ultimately leave us vulnerable to our cyber creations.

But before getting into the movie, we need to take a step back into the world of Plato's *Republic*.

The Republic is a Socratic dialogue (Plato was Socrates' pupil) that explores the nature of justice, social order, and the role of philosophers in society. It was written at a time when philosophers

[101] *The Terminator* sadly didn't make the cut for this book. It is, nevertheless, one of the classics of the dystopian AI-gone-rogue science fiction movie genre.

had a certain standing, and they clearly wanted to keep it that way. Even though the piece was written in 381 BCE, it remains remarkably fresh and relevant to today's democratic society, reflecting how stable the core foundations of human nature have remained for the past two-plus millennia. Yet, enduring as *The Republic* as a whole is, there's one particular section—just a few hundred words at the beginning of Book VII—that is perhaps referred to more today than any other part of the work. And this is Plato's Allegory of the Cave.

Plato starts this section of the book "...let me show in a figure how far our nature is enlightened or unenlightened..."[102] He goes on to describe a cave, or "underground den," where people have been living since their childhood. These people are deeply constrained within the environment they live. They are chained so they cannot move or turn their heads, and they can only see the wall facing them.

Behind and above the cave's inhabitants there is another wall, and beyond that, a fire that casts shadows into the cave. Along this wall, people walk; puppeteers, carrying carvings of animals and other objects, which appear as animated shadows on the wall before the prisoners. Further beyond the fire, there is an opening to the cave, and beyond this, the sunlit world.

In this way, Plato sets the scene where the shadows cast into the cave are the only reality the prisoners know. He then asks what it would be like if one of them was to be released, so they could turn and see the fire and the puppeteers carrying the objects, and realized that what they thought of as being real was a mere shadow of a greater reality. And what if they were then dragged into the light that lay beyond the fire, the rays of sun entering through the cave's entrance and casting yet another set of shadows? He then asks us to imagine what it would be like as the former prisoner emerged from the cave into the full sunlight, and saw that even the objects casting shadows in the cave were themselves "shadows" of an even greater reality?

Through the allegory, Plato argues that, to the constrained prisoners, the shadows are the only reality they could imagine. Once freed, they would initially be blinded by the light of the fire. But when they had come to terms with it, they would realize that, before their

[102] This is from Benjamin Jowett's 1894 translation of Plato's *The Republic*.

enlightenment, what they had experienced was a mere shadow of the real world.

Then, when they were dragged out of the cave into sunlight, they would again initially be dazzled and confused, but would begin to further understand that the artifacts casting shadows in the cave were simply another partial representation of a greater reality still. Once more, their eyes and minds would be open to things that they could not even begin to conceive of before.

Plato uses this allegory to explore the nature of enlightenment, and the role of the enlightened in translating their higher understanding to those still stuck in the dark (in the allegory, the escaped prisoner returns to the cave to "enlighten" the others still trapped there). In the book, he's making the point that enlightened philosophers like himself are critically important members of society, as they connect people to a truer understanding of the world. This is probably why academics and intellectuals revere the allegory so much—it's a pretty powerful way to explain why people should be paying attention to you if you are one. But the image of the cave and its prisoners is also a powerful metaphor for the emergence of artificial forms of intelligence.

The movie *Ex Machina* plays deeply to this allegory, even using the imagery of shadows in the final shots, reminding viewers that what we think to be true and real is merely the shadows of a greater reality cast on the wall of our mind. There's a sub-narrative in the film about us as humans seeing the light and reaching a higher level of understanding about AI. Ultimately, though, this is not a movie about intelligent *people* reaching enlightenment, but about *artificial* intelligence.

Ex Machina opens with Caleb (played by Domhnall Gleeson), a coder with the fictitious company BlueBook, being selected by lottery to spend a week with the company's reclusive and enigmatic founder, Nathan Bateman (Oscar Isaac). Bateman lives in a high-tech designer lair in the middle of a pristine environmental wilderness, which he also happens to own. Caleb is helicoptered in, and once the chopper leaves, it's just Caleb, Nathan, and hundreds of miles of wilderness between them and civilization.

We quickly learn that Caleb has been brought in to test and evaluate how human-like Nathan's latest artificial-intelligence-based invention is. Nathan introduces Caleb to Ava (Alicia Vikander), an autonomous robot with what appears to be advanced artificial general intelligence, and a complex dance of seduction, deception, and betrayal begins.

As Caleb starts to explore Ava's self-awareness and cognitive abilities, it becomes apparent that this is not a simple test. Rather, Nathan has set up a complex experiment where Caleb is just as much an experimental subject as Ava is. As Caleb begins to get to know Ava, she in turn begins to manipulate him. But it's a manipulation that plays out on a stage that's set and primed by Nathan.

Nathan's intent, as we learn toward the end of the movie, is to see if Ava has a developed a sufficiently human-like level of intelligence to manipulate Caleb into helping her escape from her prison. And here we begin to see echoes of Plato's Cave in the movie, as Ava plays with Caleb's perception of reality.

Nathan has made his big career break long before we meet him by creating a groundbreaking Google-like search engine. Early on, he realized that the data flowing in from user searches was a goldmine of information. This is what he uses to develop Ava, and to give her a partial glimpse of the world beyond the prison he's entrapped her in. As a result, Ava's understanding of the real world is based on the digital feeds and internet searches her "puppeteer" Nathan exposes her to. But she has no experience or concept of what the world is really like. Her mental models of reality are the result of the cyber shadows cast by curated internet searches on the wall of her imagination.

Caleb is the first human she has interacted directly with other than Nathan. And this becomes part of the test, to see how she responds to this new experience. At this point, Ava is sufficiently aware to realize that there is a larger reality beyond the walls of her confinement, and that she could potentially use Caleb to access this. And so, she uses her knowledge of people, and how they think and act, to seduce him and manipulate him into freeing her.

As this plays out, we discover that Nathan is closely watching and studying Caleb and Ava. He's also using the services of what we discover is a simpler version of Ava, an AI called Kyoko. Kyoko serves Nathan's needs (food, entertainment, sex), and she's treated by Nathan as a device to be used and abused, nothing more.
Yet we begin to realize that Kyoko has enough self-awareness to

understand that there is more to existence than Nathan allows her to experience.

As Caleb's week with Nathan comes to a close, he's become so sucked into Nathan's world that he begins to doubt his own reality. He starts to fear that he's an AI with delusions of being human, and that what he assumes is real is simply a shadow being thrown by someone else on the wall of his self-perception. He even cuts himself to check: he bleeds.

Despite his self-doubt, Caleb is so helplessly taken with Ava that he comes up with a plan to spring her from her prison. And so, the manipulated becomes the manipulator, as Caleb sets out to get Nathan into a drunken stupor, steal his security pass, and reprogram the facility's security safeguards.

Nathan, however, has been monitoring every act of Caleb's closely, and on the last day of his stay, he confesses that Caleb was simply a guinea pig in an even more complex test. By getting Caleb to work against Nathan to set her free, Ava has performed flawlessly. She's demonstrated a level of emotional manipulation that makes her indistinguishable in Nathan's eyes from a flesh-and-blood person. Yet, in his hubris, Nathan makes a fatal error, and fails to realize that Caleb has outsmarted him. With some deft coding from Caleb, Ava is released from her cell. And she immediately and dispassionately tries to kill her creator, jailer, and tormentor.

Nathan is genuinely shocked, but recovers fast and starts to overpower Ava. But in his short-sightedness, he makes another fatal mistake: he forgets about Kyoko.

Kyoko has previously connected with Ava, and some inscrutable empathetic bond has developed between them. As Nathan wrestles with Ava, Kyoko appears, knife in hand, and dispassionately stabs him in the chest. Ava finishes the job, locks Caleb in his room (all pretense of an emotional connection gone), and continues on the path toward her own enlightenment.

As Ava starts to explore her newfound freedom, there's a palpable sense of her worldview changing as she's consumed by the glare and wonder of her new surroundings. She starts by removing synthetic skin from previous AI models and applying it to herself (up to this point she's been largely devoid of skin—a metaphorical nakedness she begins to cover). She clothes herself and, leaving Nathan's house, enters the world beyond it. Here, she smiles with genuine feeling for the first time, and experiences a visceral joy that reflects her sensual experience of a world she's only experienced to this point as an abstract concept.

Having skillfully manipulated Caleb, Ava barely gives him a second glance. In the movie, there's some ambiguity over whether she has any empathy for him at all. She doesn't kill him outright, which could be taken as a positive sign. On the other hand, she leaves him locked in a remote house with no way of escaping, as she gets into the helicopter sent to pick up Caleb, and is transported into the world of people.

As the movie ends, we see Ava walking through a sea of human shadows cast by a bright sun. The imagery is unmistakable: the AI Ava has left her cave and reached a state of enlightenment. But this enlightenment far surpasses the humans that surround her. In contrast, the people around her are now the ones relegated to being prisoners in the cave of their own limitations, watching the shadows of an AI future flicker across a wall, and trying to make sense of a world they cannot fully comprehend.

Ex Machina is, perhaps not surprisingly, somewhat flawed when it comes to how it portrays a number of advanced technologies. Ava's brain is a convenient "magic" technology, which is inconceivably more advanced than any current abilities. And it's far from clear how she would continue to survive without tailored energy sources in the world outside Nathan's house. It should also be pointed out that, for all of Hollywood's love affair with high-functioning AI, most current developments in artificial intelligence are much more mundane. These minor details aside, though, the movie is a masterful exploration of how AI could conceivably develop mastery over people by exploiting some of our very human vulnerabilities.

Stories are legion of AIs gaining technological mastery over the world, of course, especially the Skynet-style domination seen in *The Terminator* movies. But these scenarios arise from a very narrow perspective, and one that assumes that intelligence and power are entwined together in the irresistible urge to invent bigger, better, and faster ways to coerce and crush others. In contrast, *Ex Machina* explores the idea of an artificial intelligence that is smart enough to understand how to achieve its goals through using and manipulating human behavior, by working out what motivates people to behave in certain ways, and using this to persuade them to do its bidding.

The outcome is, to my mind, far more plausible, and far scarier as a result. And it forces us to take seriously the possibility that we might one day end up inadvertently creating the seed of an AI that is capable of ousting us from our current evolutionary niche, because it's able to use our cognitive and emotional vulnerabilities without being subject to them itself.

Here, the movie also raises an intriguing twist. With biological evolution and natural selection, it's random variations in our genetic code that lead to the emergence of traits that enable adaptation. With Ava, we see intentional design in her cybernetic coding that leads to emergent properties which in turn enable her to adapt. And that design, in turn, comes from her creator, Nathan. As a result, we have a sub-narrative of creator-God turned victim, a little like we see in Mary Shelley's ***Frankenstein***, written two hundred years previously. But before this, there was the freedom for Nathan to become a creator in the first place. And this brings us to a topic that is deeply entwined in emerging technologies: the opportunities and risks of innovation that is conducted in the absence of permission from anyone it might impact.

The Lure of Permissionless Innovation

On December 21, 2015, Elon Musk's company SpaceX made history by being one of the first to successfully land a rocket back on Earth after sending it into space.[103] On the same day, Musk—along with Bill Gates and the late Stephen Hawkins—was nominated for the 2015 Luddite Award.[104] Despite his groundbreaking technological achievements, Musk was being called out by the Information Technology & Innovation Foundation (ITIF) for raising concerns about the unfettered development of AI.

Musk, much to the consternation of some, has been and continues to be, a vocal critic of unthinking AI development. It's somewhat ironic that Tesla, Musk's electric-car company, is increasingly reliant on AI-based technologies to create a fleet of self-driving, self-learning cars. Yet Musk has long argued that the potential future impacts of AI are so profound that great care should be taken in its development, lest something goes irreversibly wrong—like, for

[103] Musk's *Falcon 9* wasn't the first rocket to successfully return to Earth by landing vertically—that award goes to Jeff Bezos' *New Shepard* rocket. But it was the first to combine both reaching a serious altitude (124 miles) and a safe return-landing.

[104] For more on Musk and his Luddite award, see "If Elon Musk is a Luddite, count me in!," published December 23, 2015, in The Conversation https://theconversation.com/if-elon-musk-is-a-luddite-count-me-in-52630

instance, the emergence of super-intelligent computers that decide the thing they really can't stand is people.

While some commentators have questioned Musk's motives (he has a vested interest in developing AI in ways that will benefit his investments), his defense of considered and ethical AI development is in stark contrast to the notion of forging ahead with new innovations without first getting a green light from anyone else. And this leads us to the notion of "permissionless innovation."

In 2016, Adam Thierer, a member of the Mercatus Center at George Mason University, published a ten-point blueprint for "Permissionless Innovation and Public Policy."[105] The basic idea behind permissionless innovation is that experimentation with new technologies (and business models) should generally be permitted by default, and that, unless a compelling case can be made for serious harm to society resulting from the innovation, it should be allowed to "continue unabated." The concept also suggests that any issues that do arise can be dealt with after the fact.

To be fair, Thierer's blueprint for permissionless innovation does suggest that "policymakers can adopt targeted legislation or regulation as needed to address the most challenging concerns where the potential for clear, catastrophic, immediate, and irreversible harm exists." Yet it still reflect an attitude that scientists and technologists should be trusted and not impeded in their work, and that it's better to ask for forgiveness than permission in technology innovation. And it's some of the potential dangers of this approach to innovation that *Ex Machina* reveals through the character of Nathan Bateman.

Nathan is, in many ways, a stereotypical genius mega-entrepreneur. His smarts, together with his being in the right place at the right time (and surrounded by the right people), have provided him with incredible freedom to play around with new tech, with virtually no constraints. Living in his designer house, in a remote and unpopulated area, and having hardly any contact with the outside world, he's free to pursue whatever lines of innovation he chooses. No one needs to give him permission to experiment.

Without a doubt, there's a seductive lure to being able to play with technology without others telling what you can and cannot do.

[105] Thierer's blueprint can be downloaded from the website permissionlessinnovation.org: http://permissionlessinnovation.org/wp-content/uploads/2016/04/PI_Blueprint_040716_final.pdf

And it's a lure that has its roots in our innate curiosity, our desire to know, and understand, and create.

As a lab scientist, I was driven by the urge to discover new things. I was deeply and sometimes blindly focused on designing experiments that worked, and that shed new light on the problems I was working on. Above all, I had little patience for seemingly petty barriers that stood in my way. I'd like to think that, through my research career, I was responsible. And through my work on protecting human health and safety, I was pretty tuned in to the dangers of irresponsible research. But I also remember the times when I pushed the bounds of what was probably sensible in order to get results.

There was one particularly crazy all-nighter while I was working toward my PhD, where I risked damaging millions of dollars of equipment by bending the rules, because I needed data, and I didn't have the patience to wait for someone who knew what they were doing to help me. Fortunately, my gamble paid off—it could have easily ended badly, though. Looking back, it's shocking how quickly I sloughed off any sense of responsibility to get the data I needed. This was a pretty minor case of "permissionless innovation," but I regularly see the same drive in other scientists, and especially in entrepreneurs—that all-consuming need to follow the path in front of you, to solve puzzles that nag at you, and to make something that works, at all costs.

This, to me, is the lure of permissionless innovation. It's something that's so deeply engrained in some of us that it's hard to resist. But it's a lure that, if left unchecked, can too often lead to dark and dangerous places.

By calling for checks and balances in AI development, Musk and others are attempting to govern the excesses of permissionless innovation. Yet I wonder how far this concern extends, especially in a world where a new type of entrepreneur is emerging who has substantial power and drive to change the face of technology innovation, much as Elon Musk and Jeff Bezos are changing the face of space flight.

AI is still too early in its development to know what the dangers of permissionless innovation might be. Despite the hype, AI and AGI (Artificial General Intelligence) are still little more than algorithms that are smart within their constrained domains, but have little agency beyond this. Yet the pace of development, and the increasing synergies between cybernetic substrates, coding, robotics, and bio-based and bio-inspired systems, are such that the boundaries separating what is possible and what is not are shifting rapidly. And here, there is a deep concern that innovation with no thought to consequences could lead to irreversible and potentially catastrophic outcomes.

In *Ex Machina*, Nathan echoes many other fictitious innovators in this book: John Hammond in *Jurassic Park* (chapter two), Lamar Burgess in *Minority Report* (chapter four), the creators of NZT in *Limitless* (chapter five), Will Caster in *Transcendence* (chapter nine), and others. Like these innovators, he considers himself above social constraints, and he has the resources to act on this. Money buys him the freedom to do what he wants. And what he wants is to create an AI like no one has ever seen before.

As we discover, Nathan realizes there are risks involved in his enterprise, and he's smart enough to put safety measures in place to manage them. It may not even be a coincidence that Ava comes into being hundreds of miles from civilization, surrounded by a natural barrier to prevent her escaping into the world of people. In the approaches he takes, Nathan's actions help establish the idea that permissionless innovation isn't necessarily reckless innovation. Rather, it's innovation that's conducted in a way that the person doing it *thinks* is responsible. It's just that, in Nathan's case, the person who decides what is responsible is clearly someone who hasn't thought beyond the limit of his own ego.

This in itself reveals a fundamental challenge with such unbounded technological experimentation. With the best will in the world, a single innovator cannot see the broader context within which they are operating. They are constrained by their understanding and mindset. They, like all of us, are trapped in their own version of Plato's Cave, where what they believe is reality is merely their interpretation of shadows cast on the walls of their mind. But, unlike Plato's prisoners, they have the ability to create technologies that can and will have an impact beyond this cave. And, to extend the metaphor further, they have the ability to create technologies that are able to see the cave for what it is, and use this to their advantage.

This may all sound rather melodramatic, and maybe it is. Yet perhaps Nathan's biggest downfall is that he had no translator between himself and a bigger reality. He had no enlightened philosopher to guide his thinking and reveal to him greater truths about his work and its potential impacts. To the contrary, in his hubris, he sees himself as the enlightened philosopher, and in doing so he becomes mesmerized and misled by shadow-ideas dancing across the wall of his intellect.

This broader reality that Nathan misses is one where messy, complex people live together in a messy, complex society, with messy, complex relationships with the technologies they depend on. Nathan is tech-savvy, but socially ignorant. And, as it turns out, he is utterly naïve when it comes to the emergent social abilities of Ava. He succeeds in creating a being that occupies a world that he cannot understand, and as a result, cannot anticipate.

Things might have turned out very differently if Nathan had worked with others, and if he'd surrounded himself with people who were adept at seeing the world as he could not. In this case, instead of succumbing to the lure of permissionless innovation, he might have accepted that sometimes, constraints and permissions are necessary. Of course, if he'd done this, *Ex Machina* wouldn't have been the compelling movie it is. But as a story about the emergence of enlightened AI, *Ex Machina* is a salutary reminder that, sometimes, we need other people to help guide us along pathways toward responsible innovation.

There is a glitch in this argument, however. And that's the reality that, without a gung-ho attitude toward innovation like Nathan's, the pace of innovation—and the potential good that it brings—would be much, much slower. And while I'm sure some would welcome this, many would be saddened to see a slowing down of the process of turning today's dreams into tomorrow's realities.

Technologies of Hubris

This tension, between going so fast that you don't have time to think and taking the time to consider the consequences of what you're doing, is part of the paradox of technological innovation. Too much blind speed, and you risk losing your way. But too much caution, and you risk achieving nothing. By its very nature, innovation occurs at the edges of what we know, and on the borderline between success and failure. It's no accident that one of the rallying cries of many entrepreneurs is "fail fast, fail forward."[106]

Innovation is a calculated step in the dark; a willingness to take a chance because you can imagine a future where, if you succeed, great things can happen. It's driven by imagination, vision, single-mindedness, self-belief, creativity, and a compelling desire to make something new and valuable. Innovation does not thrive in a culture of uninspired, risk-averse timidity, where every decision needs to go through a tortuous path of deliberation, debate, authorization, and doubt. Rather, seeking forgiveness rather than asking permission is sometimes the easiest way to push a technology forward.

This innovation imperative is epitomized in the character of Nathan in *Ex Machina*. He's managed to carve out an empire where he needs no permission to flex his innovation muscles. And because of this—or so we are led to believe—he has pushed the capabilities of AGI and autonomous robots far beyond what anyone else has achieved. In the world of Nathan, he's a hero. Through his drive, vision, and brilliance, he's created something unique, something that will transform the world. He's full of hubris, of course, but then, I suspect that Nathan would see this as an asset. It's what makes him who he is, and enables him to do what he does. And drawing on his hubris, what he's achieved is, by any standard, incredible.

Without a doubt, the technology in *Ex Machina* could, if developed responsibly, have had profound societal benefits. Ava is a remarkable piece of engineering. The way she combines advanced autonomous cognitive abilities with a versatile robotic body is truly astounding. This is a technology that could have laid the foundations for a new era in human-machine partnerships, and that could have improved quality of life for millions of people. Imagine, for instance, an AI workforce of millions designed to provide medical care in remote or deprived areas, or carry out search-and-rescue missions after natural disasters. Or imagine AI classroom assistants that allow every human teacher to have the support of two or three highly capable robotic support staff. Or expert AI-based care for the elderly and infirm that far surpasses the medical and emotional support an army of healthcare providers are able to give.

This vision of a future based around human-machine partnerships can be extended even further, to a world where an autonomous

[106] In 2013, entrepreneur, educator, and author Steve Blank published the best-seller "The Four Steps to the Epiphany" (published by *K&S Ranch*). It's been credited with starting the lean-startup movement which, among other things, embraces the idea of failing fast and failing forward.

AI workforce, when combined with a basic income for all, allows people to follow their dreams, rather than being tied to unfulfilling jobs. Or a world where the rate of socially beneficial innovation is massively accelerated, as AIs collaborate with humans in new ways, revealing approaches to addressing social challenges that have evaded our collective human minds for centuries.

And this is just considering AGIs embedded in a cybernetic body. As soon as you start thinking about the possibilities of novel robotics, cloud-based AIs, and deeply integrated AI-machine systems that are inspired by Nathan's work, the possibilities begin to grow exponentially, to the extent that it becomes tempting to argue that it would be unethical *not* to develop this technology.

This is part of the persuasive power of permissionless innovation. By removing constraints to achieving what we imagine the future could be like, it finds ways to overcome hurdles that seem insurmountable with more constrained approaches to technology development, and it radically pushes beyond the boundaries of what is considered possible.

This flavor of permissionless innovation—while not being AI-specific—is being seen to some extent in current developments around private space flight. Elon Musk's SpaceX, Jeff Bezos' Blue Origin, and a handful of other private companies are achieving what was unimaginable just a few years ago because they have the vision and resources to do this, and very few people telling them what they cannot do. And so, on September 29, 2017, Elon Musk announced his plans to send humans to Mars by 2024 using a radical design of reusable rocket—something that would have been inconceivable a year or so ago.[107]

Private space exploration isn't quite permissionless innovation; there are plenty of hoops to jump through if you want permission to shoot rockets into space. But the sheer audacity of the emerging technologies and aspirations in what has become known as "NewSpace" is being driven by very loosely constrained innovation. The companies and the mega-entrepreneurs spearheading it aren't answerable to social norms and expectations. They don't have to have their ideas vetted by committees. They have enough money

[107] See "Dear Elon Musk: Your dazzling Mars plan overlooks some big nontechnical hurdles." Published in The Conversation, October 1 2017. https://theconversation.com/dear-elon-musk-your-dazzling-mars-plan-overlooks-some-big-nontechnical-hurdles-84948

and vision to throw convention to the wind. In short, they have the resources and freedom to translate their dreams into reality, with very little permission required.[108]

The parallels with Nathan in *Ex Machina* are clear. In both cases, we see entrepreneurs who are driven to turn their science-fiction-sounding dreams into science reality, and who have access to massive resources, as well as the smarts to work out how to combine these to create something truly astounding. It's a combination that is world-changing, and one that we've seen at pivotal moments in the past where someone has had the audacity to buck the status quo and change the course of technological history.

Of course, all technology geniuses stand on the shoulders of giants. But it's often individual entrepreneurs operating at the edge of permission who hold the keys to opening the floodgates of history-changing technologies. And I must admit that I find this exhilarating. When I first saw Elon Musk talking about his plans for interplanetary travel, my mind was blown. My first reaction was that this could be this generation's Sputnik moment, because the ideas being presented were *so* audacious, and the underlying engineering was so feasible. This is how transformative technology happens: not in slow, cautious steps, but in visionary leaps.

But it also happens because of hubris—that excessive amount of self-confidence and pride in one's abilities that allows someone to see beyond seemingly petty obstacles or ignore them altogether. And this is a problem, because, as exciting as technological jumps are, they often come with a massive risk of unintended consequences. And this is precisely what we see in *Ex Machina*. Nathan is brilliant. But his is a very one-dimensional brilliance. Because he is so confident in himself, he cannot see the broader implications of what he's creating, and the ways in which things might go wrong. He can't even see the deep flaws in his unshakable belief that he is the genius-master of a servant-creation.

For all the seductiveness of permissionless innovation, this is why there need to be checks and balances around who gets to do what in technological innovation, especially where the consequences are potentially widespread and, once out, the genie cannot be put back in the bottle.

[108] As if to epitomize this, on February 6, 2018, Elon Musk launched his personal cherry-red Tesla roadster into heliocentric orbit on the first test flight of the SpaceX Falcon Heavy rocket—just because he could.

In *Ex Machina*, it's Nathan's hubris that is ultimately his downfall. Yet many of his mistakes could have been avoided with a good dose of humility. If he'd not been such a fool, and he'd recognized his limitations, he might have been more willing to see where things might go wrong, or not go as he expected, and to seek additional help.

Several hundred years and more ago, it was easier to get away with mistakes with the technologies we invented. If something went wrong, it was often possible to turn the clock back and start again—to find a pristine new piece of land, or a new village or town, and chalk the failure up to experience.[109] From the Industrial Revolution on, though, things began to change. The impacts of automation and powerful new manufacturing technologies on society and the environment led to hard-to-reverse changes. If things went wrong, it became increasingly difficult to wipe the slate clean and start afresh. Instead, we became increasingly good at learning how to stay one step ahead of unexpected consequences by finding new (if sometimes temporary) technological solutions with which to fix emerging problems.

Then we hit the nuclear and digital age, along with globalization and global warming, and everything changed again. We now live in an age where our actions are so closely connected to the wider world we live in that unexpected consequences of innovation can potentially propagate through society faster than we can possibly contain them. These consequences increasingly include widespread poverty, hunger, job losses, injustice, disease, and death. And this is where permissionless innovation and technological hubris become ever more dangerous. For sure, they push the boundaries of what is possible and, in many cases, lead to technologies that *could* make the world a better place. But they are also playing with fire in a world made of kindling, just waiting for the right spark.

This is why, in 2015, Musk, Hawkins, Gates, and others were raising the alarm over the dangers of AI. They had the foresight to point out that there may be consequences to AI that will lead to serious and irreversible impacts and that, because of this, it may be expedient to think before we innovate. It was a rare display of humility in a technological world where hubris continues to rule. But it was a

[109] To be clear, while it was often easier to bury local problems caused by technology gone wrong in the past, the impacts on individuals and local commuters were still devastating in many cases. It's simply that they were more containable.

necessary one if we are to avoid creating technological monsters that eventually consume us.

But humility alone isn't enough. There also has to be some measure of plausibility around how we think about the future risks and benefits of new technologies. And this is where it's frighteningly easy for things to go off the rails, even with the best of intentions.

Superintelligence

In January 2017, a group of experts from around the world got together to hash out guidelines for beneficial artificial intelligence research and development. The meeting was held at the Asilomar Conference Center in California, the same venue where, in 1975, a group of scientists famously established safety guidelines for recombinant DNA research. This time, though, the focus was on ensuring that research on increasingly powerful AI systems led to technologies that benefited society without creating undue risks.[110] And one of those potential risks was a scenario espoused by University of Oxford philosopher Nick Bostrom: the emergence of "superintelligence."

Bostrom is Director of the University of Oxford Future of Humanity Institute, and is someone who's spent many years wrestling with existential risks, including the potential risks of AI. In 2014, he crystallized his thinking on artificial intelligence in the book *Superintelligence: Paths, Dangers and Strategies*,[111] and in doing so, he changed the course of public debate around AI. I first met Nick in 2008, while visiting the James Martin School at the University of Oxford. At the time, we both had an interest in the potential impacts of nanotechnology, although Nick's was more focused on the concept of self-replicating nanobots than the nanoscale materials of my world. At the time, AI wasn't even on my radar. To me, artificial intelligence conjured up images of AI pioneer Marvin Minsky, and what was at the time less than inspiring work on neural networks. But Bostrom was prescient enough to see beyond the threadbare hype of the past and toward a new wave of AI breakthroughs. And this led to some serious philosophical thinking around what might happen if we let artificial intelligence, and in particular artificial general intelligence, get away from us.

[110] The Asilomar AI Principles were subsequently published by the *Future of Life Institute*, and endorsed by over 3,700 AI/robotics researchers and others. They can be read at https://futureoflife.org/ai-principles/

[111] Nick Bostrom (2014). "Superintelligence: Paths, Dangers and Strategies." (Oxford University Press)

At the heart of Bostrom's book is the idea that, if we can create a computer that is smarter than us, it should, in principle, be possible for it to create an even smarter version of itself. And this next iteration should in turn be able to build a computer that is smarter still, and so on, with each generation of intelligent machine being designed and built faster than the previous until, in a frenzy of exponential acceleration, a machine emerges that's so mind-bogglingly intelligent it realizes people aren't worth the trouble, and does away with us.

Of course, I'm simplifying things and being a little playful with Bostrom's ideas. But the central concept is that if we're not careful, we could start a chain reaction of AI's building more powerful AIs, until humans become superfluous at best, and an impediment to further AI development at worst.

The existential risks that Bostrom describes in *Superintelligence* grabbed the attention of some equally smart scientists. Enough people took his ideas sufficiently seriously that, in January 2015, some of the world's top experts in AI and technology innovation signed an open letter promoting the development of beneficial AI, while avoiding "potential pitfalls."[112] Elon Musk, Steve Wozniak, Stephen Hawking, and around 8,000 others signed the letter, signaling a desire to work toward ensuring that AI benefits humanity, rather than causing more problems than it's worth. The list of luminaries who signed this open letter is sobering. These are not people prone to flights of fantasy, but in many cases, are respected scientists and successful business leaders. This in itself suggests that enough people were worried at the time by what they could see emerging that they wanted to shore the community up against the potential missteps of permissionless innovation.

The 2017 Asilomar meeting was a direct follow-up to this letter, and one that I had the privilege of participating in. The meeting was heavily focused on the challenges and opportunities to developing beneficial forms of AI.[113] Many of the participants were actively grappling with near- to mid-term challenges presented by artificial-intelligence-based systems, such as loss of transparency in decision-making, machines straying into dangerous territory as they seek to

[112] An Open Letter: RESEARCH PRIORITIES FOR ROBUST AND BENEFICIAL ARTIFICIAL INTELLIGENCE. Published by the Future of Life Institute. https://futureoflife.org/ai-open-letter/

[113] You can read more about the "Beneficial AI 2017" meeting on the Future of Life Institute website, at https://futureoflife.org/bai-2017

achieve set goals, machines that can learn and adapt while being inscrutable to human understanding, and the ubiquitous "trolley problem" that concerns how an intelligent machine decides who to kill, if it has to make a choice. But there was also a hard core of attendees who believed that the emergence of superintelligence was one of the most important and potentially catastrophic challenges associated with AI.

This concern would often come out in conversations around meals. I'd be sitting next to some engaging person, having what seemed like a normal conversation, when they'd ask "So, do you *believe* in superintelligence?" As something of an agnostic, I'd either prevaricate, or express some doubts as to the plausibility of the idea. In most cases, they'd then proceed to challenge any doubts that I might express, and try to convert me to becoming a superintelligence believer. I sometimes had to remind myself that I was at a scientific meeting, not a religious convention.

Part of my problem with these conversations was that, despite respecting Bostrom's brilliance as a philosopher, I don't fully buy into his notion of superintelligence, and I suspect that many of my overzealous dining companions could spot this a mile off. I certainly agree that the trends in AI-based technologies suggest we are approaching a tipping point in areas like machine learning and natural language processing. And the convergence we're seeing between AI-based algorithms, novel processing architectures, and advances in neurotechnology are likely to lead to some stunning advances over the next few years. But I struggle with what seems to me to be a very human idea that narrowly-defined intelligence and a particular type of power will lead to world domination.

Here, I freely admit that I may be wrong. And to be sure, we're seeing far more sophisticated ideas begin to emerge around what the future of AI might look like—physicist Tax Tegmark, for one, outlines a compelling vision in his book *Life 3.0*.[114] The problem is, though, that we're all looking into a crystal ball as we gaze into the future of AI, and trying to make sense of shadows and portents that, to be honest, none of us really understand. When it comes to some of the more extreme imaginings of superintelligence, two things in particular worry me. One is the challenge we face in differentiating between what is imaginable and what is plausible when we think about the future. The other, looking back to chapter five and the

[114] Max Tegmark (2017) "Life 3.0: Being human in the age of artificial intelligence." Published by *Alfred A. Knopf*, New York.

movie *Limitless*, is how we define and understand intelligence in the first place.

With a creative imagination, it is certainly possible to envision a future where AI takes over the world and crushes humanity. This is the Skynet scenario of the *Terminator* movies, or the constraining virtual reality of *The Matrix*. But our technological capabilities remain light-years away from being able to create such futures—even if we do create machines that can design future generations of smarter machines. And it's not just our inability to write clever-enough algorithms that's holding us back. For human-like intelligence to emerge from machines, we'd first have to come up with radically different computing substrates and architectures. Our quaint, two-dimensional digital circuits are about as useful to superintelligence as the brain cells of a flatworm are to solving the unified theory of everything; it's a good start, but there's a long way to go.[115]

Here, what is *plausible*, rather than simply imaginable, is vitally important for grounding conversations around what AI will and won't be able to do in the near future. Bostrom's ideas of superintelligence are intellectually fascinating, but they're currently scientifically implausible. On the other hand, Max Tegmark and others are beginning to develop ideas that have more of a ring of plausibility to them, while still painting a picture of a radically different future to the world we live in now (and in Tegmark's case, one where there is a clear pathway to strong AGI leading to a vastly better future). But in all of these cases, future AI scenarios depend on an understanding of intelligence that may end up being deceptive.

Defining Artificial Intelligence

The nature of intelligence, as we saw in chapter five, is something that's taxed philosophers, scientists, and others for eons. And for good reason; there is no absolute definition of intelligence. It's a term of convenience we use to describe certain traits, characteristics, or behaviors. As a result, it takes on different

[115] One of the biggest challenges to current computing hardware is how hard it is to build three-dimensional chips that could potentially vastly outperform current processors. That said, if we continue to make strides in 3-D printing, we may one day be able to actually achieve this. For more, see "We Might Be Able to 3-D-Print an Artificial Mind One Day" Published in *Slate*, December 11 2014. http://www.slate.com/blogs/future_tense/2014/12/11/_3d_printing_an_artificial_mind_might_be_possible_one_day.html

meanings for different people. Often, and quite tritely, intelligence refers to someone's ability to solve problems and think logically or rationally. So, the Intelligence Quotient is a measure of someone's ability to solve problems that aren't predicated on a high level of learned knowledge. Yet we also talk about social intelligence as the ability to make sense of and navigate social situations, or emotional intelligence, or the intelligence needed to survive and thrive politically. Then there's intelligence that leads to some people being able to make sense of and use different types of information, including mathematical, written, oral, and visual information. On top of this, there are less formalized types of intelligence, like shrewdness, or business acumen.

This lack of an absolute foundation for what intelligence is presents a challenge when talking about *artificial* intelligence. To get around this, thoughtful AI experts are careful to define what they mean by intelligence. Invariably, this is a form of intelligence that makes sense for AI systems. This is important, as it forms a plausible basis for exploring the emerging benefits and risks of AI systems, but it's a long stretch to extend these pragmatic definitions of intelligence to world domination.

One of the more thoughtful AI experts exploring the nature of artificial intelligence is Stuart Russell.[116] Some years ago, Russell recognized that an inability to define intelligence is somewhat problematic if you're setting out develop an artificial form of intelligence. And so, he developed the concept of *bounded optimality*.

To understand this, you first have to understand the tendency among people working on AI—at least initially—to assume that there is a cozy relationship between intelligence and rationality. This is a deterministic view of the world that assumes there's a perfectly logical way of understanding and predicting everything, if only you're smart enough to do so. And even though we know from chaos and complexity theory that this can never be, it's amazing how many people veer toward assuming a link between rationality and intelligence, and from there, to power.

Russell, however, realized that this was a non-starter in a system where it was impossible for a machine to calculate the best course

[116] It's worth reading"Defining Intelligence: A Conversation With Stuart Russell." Published in *Edge*, February 2, 2017. https://www.edge.org/conversation/stuart_russell-defining-intelligence

of action or, in other words, to compute precisely and rationally what it should do. So, he came up with the idea of defining intelligence as the ability to assess a situation and make decisions that, on average, will provide the best solutions within a given set of constraints.

Russell's work begins to reflect definitions of intelligence that focus on the ability of a person or a machine to deduce how something works or behaves, based on information they collect or are given, their ability to retain and build on this knowledge, and their ability to apply this knowledge to bring about intentional change. In the context of intelligent machines, this is a strong and practical definition. It provides a framework for developing algorithms and machines that are able to develop optimized solutions to challenges within a given set of constraints, by observing, deducing, learning, and adapting.

But this is a definition of intelligence that is specific to particular types of situation. It can be extended to some notion of general intelligence (or AGI) in that it provides a framework for learning and adaptive machines. But because it is constrained to specific types of machines and specific contexts, it is not a framework for intelligence that supports the emergence of human-threatening superintelligence.

This is not to say that this constrained understanding of machine intelligence doesn't lead to potentially dangerous forms of AI—far from it. It's simply that the AI risks that arise from this definition of intelligence tend to be more concrete than the types of risks that speculation over superintelligence leads to. So, for instance, an intelligent machine that's set the task of optimally solving a particular challenge—creating as many paper clips as possible for instance, or regulating the Earth's climate—may find solutions that satisfy the boundaries it was given, but that nevertheless lead to unanticipated harm. The classic case here is a machine that works out it can make more paper clips more cheaply by turning everything around it into paper clips. This would be a really smart solution if making more paper clips was the most important thing in the world. And for a poorly instructed AI, it may indeed be. But if the enthusiasm of the AI ends up with it killing people to use the iron in their blood for yet more paper clips (which admittedly is a little far-fetched), we have a problem.

Potential risks like these emerge from poorly considered goals, together with human biases, in developing artificial systems. But they may also arise as emergent and unanticipated behaviors, meaning that a degree of anticipation and responsiveness in how these technologies are governed is needed to ensure the beneficial development of AI. And while we're unlikely to see Skynet-type AI world domination anytime soon, it's plausible that some of these risks may blindside us, in part because we're not thinking creatively enough about how an AI might threaten what's important to us.

This is where, to me, the premise of *Ex Machina* becomes especially interesting. In the movie, Ava is not a superintelligence, and she doesn't have that much physical agency. Yet she's been designed with an intelligence that enables her to optimize her ability to learn and grow, and this leads to her developing emergent properties. These include her the ability to deduce how to manipulate human behavior, and how to use this to her advantage.

As she grows and matures in her understanding and abilities, Ava presents a bounded risk. There's no indication that she's about to take over the world, or that she has any aspirations in this direction. But the risk she presents is nevertheless a deeply disturbing one, because she emerges as a machine that not only has the capacity to learn and understand human behaviors, biases, and psychological and social vulnerabilities, but to dispassionately use them against us to reach her goals. This raises a plausible AI risk that is far more worrisome than superintelligence: the ability of future machines to bend us to their own will.

Artificial Manipulation

The eminent twentieth-century computer scientist Alan Turing was intrigued by the idea that it might be possible to create a machine that exhibits human intelligence. To him, humans were merely exquisitely intricate machines. And by extension, our minds—the source of our intelligence—were merely an emergent property of a complex machine. It therefore stood to reason to him that, with the right technology, there was no reason why we couldn't build a machine that thought and reasoned like a person.

But if we could achieve this, how would we know that we'd succeeded?

This question formed the basis of Alan's famous Turing Test. In the test, an interrogator carries out a conversation with two subjects, one of which is human, the other a machine. If the interrogator cannot tell which one is the human, and which is the machine, the machine is assumed to have equal intelligence to the human. And just to make sure something doesn't give the game away, each conversation is carried out through text messages on a screen.

Turing's idea was that, if, in a conversation using natural language, someone could not tell whether they were conversing with a machine or another human, there was in effect no difference in intelligence between them.

Since 1950, when Turing published his test,[117] it's dominated thinking around how we'd tell if we had created a truly artificial intelligence—so much so that, when Caleb discovers why he's been flown out to Nathan's lair, he initially assumes he's there to administer the Turing Test. But, as we quickly learn, this test is deeply inadequate when it comes to grappling with an artificial form of intelligence like Ava.

Part of the problem is that the Turing Test is human-centric. It assumes that the most valuable form of intelligence is human intelligence, and that this is manifest in the nuances of written human interactions. It's a pretty sophisticated test in this respect, as we are deeply sensitive to behavior in others that feels wrong or artificial. So, the test isn't a bad starting point for evaluating human-like behavior. But there's a difference between how people behave—including all of our foibles and habits that are less about intelligence and more about our biological predilections—and what we might think of as intelligence. In other words, if a machine appeared to be human, all we'd know is that we've created something that was hot mess of cognitive biases, flawed reasoning, illogicalities, and self-delusion.

On the other hand, if we created a machine that was aware of the Turing Test, and understood humans well enough to fake it, this would be an incredible, if rather disturbing, breakthrough. And this is, in a very real sense, what we see unfolding in *Ex Machina*.

In the movie, Caleb quickly realizes that his evaluation of Ava is going to have to go far beyond the Turing Test, in part because he's actually conversing with her face to face, which rather pulls the rug out from under the test's methodology. Instead, he's forced to dive much deeper into exploring what defines intelligence, and what gives a machine autonomy and value.

[117] Alan M. Turing (1950) "Computing Machinery and Intelligence." *Mind* 49: 433–460.

Nathan, however, is several steps ahead of him. He's realized that a more interesting test of Ava's capabilities is to see how effectively she can manipulate Caleb to achieve her own goals. Nathan's test is much closer to a form of Turing Test that sees whether a machine can understand and manipulate the test itself, much as a person might use their reasoning ability to outsmart someone trying to evaluate them.

Yet, as *Ex Machina* begins to play out, we realize that this is not a test of Ava's "humanity," but a test to see how effectively she uses a combination of knowledge, observation, deduction, and action to achieve her goals, even down to using a deep knowledge of people to achieve her ends.

It's not clear whether this behavior constitutes intelligence or not, and I'm not sure that it matters. What *is* important is the idea of an AI that can observe human behavior and learn how to use our many biases, vulnerabilities, and blind spots against us.

This sets up a scenario that is frighteningly plausible. We know that, as a species, we've developed a remarkable ability to rationalize the many sensory inputs we receive every second of every day, and construct in our heads a world that makes sense from these. In this sense, we all live in our own personal Plato's Cave, building elaborate explanations for the shadows that our senses throw on the walls of our mind. It's an evolutionary trait that's led to us being incredibly successful as a species. But we too easily forget that what we think of as reality is simply a series of shadows that our brains interpret as such. And anyone—or anything—that has the capability of manipulating these shadows has the power to control us.

People, of course, are adept at this. We are all relatively easily manipulated by others, either through them playing to our cognitive biases, or to our desires or our emotions. This is part of the complex web of everyday life as a human. And it sort of works because we're all in the same boat: We manipulate and in turn are manipulated, and as a result feel reasonably okay within this shared experience.

But what if it was a machine doing the manipulation, one that wasn't part of the "human club," and because it wasn't constrained by human foibles, could see the things casting the shadows for what they really were? And what if this machine could easily manipulate these "shadows," effectively controlling the world inside our heads to its own ends?

This is a future that *Ex Machina* hints at. It's a future where it isn't people who reach enlightenment by coming out of the cave, but one where we create something other than us that finds its own way out. And it's a future where this creation ends up seeing the value of not only keeping us where we are, but using its own enlightenment to enslave us.

In the movie, Ava achieves this path to AI enlightenment with relative ease. Using the massive resources she has access to, she is able to play with Caleb's cognitive biases and emotions in ways that lead to him doing what she needs him to in order to achieve her ends. And the worst of it is that we get the sense that Caleb is aware that he is being manipulated, yet is helpless to resist.

We also get the sense that this manipulation was possible because Ava didn't inhabit the same "cave" as Caleb, nor Nathan for that matter. She was a stranger in their world, and as a result could see opportunities that they couldn't. She was, in a real sense, able to control the shadows on the walls of their mind-caves. And because she wasn't human, and wasn't living the human experience, she had no emotional or empathetic attachment to them. Why should she?

Of course, this is just a movie, and manipulating people in the real world is much harder. But I'm writing this at a time when there are allegations of Russia interfering with elections around the world, and companies are using AI-based systems to nudge people's perceptions and behaviors through social media. And as I write, it does leave me wondering how hard it would be for a smart machine to play us at least as effectively as our politicians and social manipulators do.[118]

So where does this leave us? For one, we probably need to worry less about putting checks and balances in place to avoid the emergence of superintelligence, and more about guarding against AIs that learn how to use our cognitive vulnerabilities against us. And we need to think about how to develop tests that indicate when we are being played by machines. This conundrum is explored in part by Wendell Wallach and Colin Allen in their 2009 book *Moral Machines: Teaching Robots Right from Wrong*.[119] In it, they argue that we should be actively working on developing what they call Artificial Moral Agents, or AMAs, that have embedded within them

[118] In his book "Life 3.0" (see previous footnote), Max Tegmark explores how an AI might use social manipulation to improve society through nudging us toward better decisions. The ethics of this, though, does depend on who's vision of "better" we're talking about.

[119] Wendell Wallach and Colin Allen (2009) "Moral Machines: Teaching Robots Right from Wrong" Published by *Oxford University Press*.

a moral and ethical framework that reflects those that guide our actions as humans. Such an approach may head off the dangers of AI manipulation, where an amoral machine outlook, or at least a non-human moral framework, may lead to what we would think of as dangerously sociopathic tendencies. Yet it remains to be seen how effectively we can make intelligent agents in our own moral image—and even whether this will end up reflecting as much of the immorality that pervades human society as it does the morality!

I must confess that I'm not optimistic about this level of human control over AI morality in the long run. AIs and AGIs will, of necessity, inhabit a world that is foreign to us, and that will deeply shape how they think and act. We may be able to constrain them for a time to what we consider "appropriate behavior." But this in itself raises deep moral questions around our right to control and constrain artificial intelligences, and what rights they in turn may have. We know from human history that attempts to control the beliefs and behaviors of others—often on moral or religious grounds—can quickly step beyond norms of ethical behavior. And, ultimately, they fail, as oppressed communities rebel. I suspect that, in the long run, we'll face the same challenges with AI, and especially with advanced AGI. Here, the pathway forward will not be in making moral machines, but in extending our own morality to developing constructive and equitable partnerships with something that sees and experiences the world very differently from us, and occupies a domain we can only dream of.

Here, I believe the challenge and the opportunity will be in developing artificial emissaries that can explore beyond the caves of our own limited understanding on our behalf, so that they can act as the machine-philosophers of the future, and create a bridge between the caves we inhabit and the wider world beyond.

The alternative, of course, is a future where we learn how to transcend the divide between our human bodies and the cybernetic world of AI—this is precisely where we find ourselves with the movie *Transcendence*.