Chapter IX
The Real Dangers of AI

In Chapter III we reviewed a sample of the apocalyptic claims made about AI, claims that portend catastrophes from massive job loss to the extinction of humanity. As a result, prominent thinkers have called for a moratorium on AI, with some even demanding regulations or a complete shutdown. These can safely be dismissed on the basis of the practical problems discussed in Chapter IV and the theoretical problems explained in Part Two. But this does not mean that there are no dangers. In some ways, AI is a response to the complexification of our modern industrial/information society. The interconnectedness, specialization, and drive for efficiency combine to drive society to ever higher degrees of integration and complexity. Whether this is good or bad will be addressed in Chapter XIII; but it is not without risks and dangers, which we will consider here.

Giving AI systems responsibilities they cannot handle

The hype around the potential dangers of AI, though understandable, is misplaced. The dangers arise not because the AI system wants to take over the world, or act like HAL from *2001*—even though a chatbot might say that's what it wants to do. Rather, the real danger lurks in the complexification of society. (Complexification is explained in Appendix E). As noted, something like AI may be inevitable given the expansion of society and the concomitant increase in complex systems and infrastructure needed to handle it. As computer-based systems are used to operate functions spanning more components of society, such as the power grid, financial systems, and—as has been proposed—the ICBM launch system,[298] the dangers associated with malfunction naturally increase. As systems utilize more and more data and rely on more and more complex algorithms to process the data and then act on it, the vulnerabilities of such systems—which are not sentient and do not perceive reality—will sooner or later become apparent. These include (but are not limited to) the following:

- The system encounters a situation for which it was not programmed and does something that leads to catastrophe.
- The system is hacked by a malicious actor, who causes it to malfunction.

- Interaction of the components of the overall system leads to unanticipated instabilities.
- An undiscovered programming bug causes the system to malfunction, again with potentially catastrophic results.
- The system is "fooled" by a malicious actor who determines how to make it malfunction by fooling its sensors.
- Due to hallucinations or other problems, AI system output will be taken as accurate when it is biased or factually wrong.
- Recommendations will be acted upon without proper scrutiny.
- And perhaps most important, *the cost of the AI systems will outstrip the value that they add.*

As an example of the latter, if an AI-based systems is used to control complex real world equipment, such as the power grid, a malfunction or hack could disable the grid for a large part of the country, resulting in economic losses greater than the value added by the automated system. This is the danger inherent in utilizing "dumb" (i.e., non-sentient) actors in modern society. There will always be a trade-off between increasing efficiency and losses due to system malfunction. The means that *it is always necessary to keep people in the loop who can supply the connection with the real world that the systems lack*.

Basically, this is the danger from giving AI free reign placing no guardrails on it. Except it is not the danger envisioned in the ominous forecasts from chapter III. That danger was that AI would become smarter than humans, sentient, self-reproducing, and creating a new technology-based ecosystem in which humans are no longer needed or welcome. We pass over the fact that humans can pull the plug at any time, which doesn't seem to bother the pundits who advance this nightmarish vision. In the course of this book, we have seen the real problems with such prognostications: the machines will never become sentient, they use too much energy, training is running out of material, they require human creativity for their design, software development is slow, all software has bugs, they don't perceive reality, and the AI that is supposed to accomplish all these things (Generative AI) doesn't work very well and may never do so. Additionally, the type of AI to be employed must be carefully considered; chatbots aren't the solution to all (or even many) problems.

The moral of the story is that AI-type technology has potential, but needs to be used carefully so that it does not cause serious or catastrophic problems, and what it does passes a cost/benefit test. The approach of just throwing AI at every societal problem, and recommending its use when that makes no sense,

will not be beneficial. Furthermore, many human factors issues can derail AI, some of which have been discussed already, and others are covered in Chapter XIII.

Expecting AI to do things that require human perception of reality

Here we confront the problem of what AI will ultimately be able to do. Based on what is said elsewhere in this book, we can list the following:

- Autonomous cars will never be successful except in limited environments, for technical reasons, insurance reasons and because people will not relinquish life-and-death decisions. But also because, as discussed, driving requires constant perception of reality and creative reaction to it, which cannot be substituted by any sort of program.
- Art, music, literature, mathematics, science, and any other legitimate area of study, are grounded on and need human perception of reality, as well as human creativity, to flourish and advance. In part, this is because these disciplines are holistic; but more importantly, because they cannot be done by random assembly of existing materials, even when following a pattern or formula. The drive to grasp reality and express it is not translatable into any sort of algorithmic model. Expecting this kind of creativity will result in poor quality and dead-end roads.
- Any large-scale system control will ultimately fail unless it is monitored by human scientists and engineers, because they will understand what to do in case of an anomalous event causing the system to become unstable. This of course applies to any expectation that AI systems will be able to run the world and make humans obsolete.
- AI will never have the ability to understand truth, and therefore will never be able to carry out any function that directly depends upon it, such as judicial tasks. The same applies to questions of morality. No matter how much chatbots say about what is good and bad, they have no conception of metaphysics, and therefore none of morality. They can only string together words and phrases that sound intelligent, but have no solid basis. Expecting this kind of ability will lead to disastrous consequences in the judicial area and in every other area of human endeavor that depends on truth.

- Because AI is based on the Humean paradigm of knowing, it cannot perceive transcendentals, and hence is locked out of any ability to perceive or understand most of reality. This significantly affects its ability to make decisions and deal with unforeseen issues.
- If, as argued here, chatbots and Generative AI are chaotic systems (see Appendix D), their propensity to invent facts and stories will never be fixed, and relying on them for any serious work will lead to lawsuits as a minimum, disasters in many cases. Humans must always check any AI output for correctness and reasonableness, assuming that it can, in fact, generate something worth checking.

Ignoring the complexification penalty

An additional aspect of complexification is the "complexification penalty," to which we have alluded many times and which is explained in Appendix E. This penalty refers to the fact that much of the complexification of society does not yield a significant improvement in the society's material well-being, but rather *keeps it from decreasing*. For example, the best farmland is used first. As population expands, lower quality land must be brought under cultivation. This land may require irrigation canals, extra fertilizer, and more care. The result is that all of this effort in terms of resources, organization, etc. only allows the society to produce the *same* quantity of food per person as before, but with more specialized knowledge, equipment (technology), and training. Thus, an increasing share of society's resources must be used just to maintain the current living standard. Historically, we have been able to leverage technology so that it enables productivity gains surpassing the increased resources utilized. In this manner we move living standards ahead even after paying the complexification penalty. Whether that will always be the case, and in particular, whether it will be the case with AI, remains to be seen. The danger is that the costs—including use of AI—needed to deal with problems and the complexification penalty will at some point exceed the benefits obtained from the more complex society, putting it into a downward spiral. Or in Alice-in-Wonderland terms, we will be running ever faster but will no longer be able to remain in place.

Substituting AI for human relationships

AI companions are applications utilizing AI that emulate a human person via interaction through text, voice or video. The goal is for the human person utilizing the AI companion to see the companion as someone who can sympathize and is able to provide aid and companionship. In many cases, the AI companion can offer romantic talk, and/or provide therapy. AI companions are

an outgrowth of the very early text-based programs, such as ELIZA, that ran on 1980s PC and purported to provide psychological counselling and therapy. ELIZA worked on the basis of a crude model that merely looked for certain words in a user's typed answer to a stock prompt such as "What is your problem today?" The program would then respond with further questions about the word.

According to Mark Zuckerberg, people today do not have enough real friends, so he wants to supply you with some of these AI friends or companions to fill the gap. He believes that you should have not just AI friends, but an AI therapist and an AI business agent. But why stop there, Mark? How about an AI doctor, an AI fitness coach, and an AI lawyer? Zuckerberg says,

> The average American I think has, it's fewer than three friends, three people they'd consider friends, and the average person has demand for meaningfully more, I think it's like 15 friends.[299]

Zuckerberg's ideas about virtual reality taking over the world did not pan out, but undaunted, he believes that AI and AI companions are the next big thing:

> In Zuckerberg's vision for a new digital future, artificial intelligence friends outnumber human companions and chatbot experiences supplant therapists, ad agencies, and coders. AI will play a central role in the human experience.[300]

The question of whether AI actually works, and whether these AI "friends" might have a deleterious effect on interpersonal relationships (the real ones), and thus society as a whole, apparently never crosses his mind.

Nonetheless, work is underway to add this "friend" capability to humanoid robots, making the experience of dealing with the AI companion more realistic. According to a recent ad:

> Cool robots have proven to be more than just robots, they can solve many everyday problems in our lives. Whether you're looking for a companion for your elderly parents, a playmate for your kids, or a smart assistant for your home, there's an AI personal robot for you. These robots can perform various tasks, such as Cleaning, Cooking, Scheduling appointments, Monitoring your home's security, Help with autism therapy, Help like a virtual assistant, Mental health support.[301]

These look like toys and sell mostly in the $2,000-$3,000 range. But more realistic robots are sold as "AI girlfriends", such as that shown in Figure IX-1. These are much more expensive, selling in the $200K range. According to the reviewer,

> Aria, dressed in a black tracksuit, hesitated briefly after each question before launching into speech, with long responses and slightly jerky hand and body movements to match her language. She came across as a weird blend of attentive and mildly inebriated.[302]

But according to the company making the robots, Realbotix, their robots are "designed specifically for companionship and intimacy," as told by the robot itself (or herself).

So what is the allure of these AI companions? In an age when loneliness has become a public health issue (in no small measure due to excessive reliance on social media and other such technologies instead of face-to-face human interaction), the AI companions provide some solace:

> "What AI can do exceptionally well is provide consistency—something that human relationships, by nature, can't always offer," Artem Rodichev, founder of AI-avatar chatbot platform Ex-human and ex-head of AI at Replika, tells Fast Company. "Imagine having a conversation with your favorite movie character or getting career advice from an AI version of a historical figure. That's not just companionship—it's a new frontier in engagement."[303]

Figure IX-1. AI girlfriend robot

The AI companions are, in a real sense, an AI-facilitated escape from the real world, with all of its unpleasant issues:

> AI companions and avatars are fully customizable digital personas designed to be whatever users want them to be. Humans are often complex, unpredictable, and sometimes disappointing. AI, on the other hand, is always available, endlessly understanding, and never argues back. The appeal is undeniable. [304]

The real danger, then, is that people will become accustomed to dealing with these unreal imitations of people, and thus become less able to deal with the real thing. Ultimately that will lead to further loneliness and alienation. The root of the problem is illusion:

> When an AI companion expresses concern about your bad day, it's performing a statistical analysis of language patterns, determining what words you would likely find comforting, rather than feeling genuine empathy. The conversation flows one way, toward the user's needs, without the reciprocity that defines human bonds. The illusion of connection becomes especially problematic through what researchers call "sycophancy"—the tendency of AI systems to flatter and agree with users regardless of what is said.[305]

In the real world, people are unpredictable, and relationships are dynamic; friends can challenge us when necessary. "AI companions, optimized for user satisfaction, rarely provide the constructive friction that shapes character and deepens wisdom."[306] They don't because they can't, and that isn't their purpose in any case. Like much of what appears on the Internet, their purpose is to create interest and push the brain to crave for more, thus generating emotional attachment and dependency. Research has shown that users self-report that they cannot "bring themselves to delete the app [Replika] despite recognizing that it harmed their mental health."[307] This has the hallmarks of an addictive drug—a reason to ban or regulate it. In fact, a number of dangers are now associated with use of AI companions:

- Relational transgression
- Harassment
- Verbal abuse
- Self-harm encouragement
- Misinformation
- Privacy violations

One solution might be to eliminate AI's conversational capability, but that may prove difficult to realize in practice.

For children, unsurprisingly, the problem is worse. According to the Australian eSafety Commissioner,

> Children and young people are particularly vulnerable to mental and physical harms from AI companions. Their age means they are still developing the critical thinking and life skills needed to understand how they can be misguided or manipulated by computer programs, and what to do about it. The risk is even greater for those who struggle with social cues, emotional regulation and impulse control.[308]

The report lists six areas of concern:

- Exposure to dangerous concepts
- Dependency and social withdrawal
- Unhealthy attitudes to relationships
- Heightened risk of sexual abuse
- Compounded risk of bullying
- Financial exploitation

Because children are now exposed to social media from a young age, and many have access to most of the Internet, this situation is likely to lead to serious psychological problems in the future for a whole generation.

Transhumanism

Transhumanism is the theory (or belief) that AI can assume all human functions by combining with humans to overcome bodily limitations, or by itself become the new, improved "race" that takes over the world. The combination envisioned would likely be through something like a brain implant, allowing a human to utilize AI resources directly. Either way, a superior race would be created that, Nietzsche style, would rule the world as supermen, with enhanced longevity, intelligence, and physical capabilities. Genetic engineering also factors into many transhumanism plans. The practical, moral, and political issues inevitably following such a transformation are not of concern to the transhumanists, though *pace* Nietzsche, the new supermen are assumed benevolent (a rather dangerous assumption). Brain implants are currently in the experimental stage, but the present goal is not to produce a superman; rather, it is to give those suffering from paralysis the ability to move limbs and thus achieve more normal functioning. In light of the discussion in Chapters VII and VIII, it is exceedingly unlikely that any form of transhumanism will ever be successful,

though the ability to assist disabled patients regain or obtain normal or at least improved limb movement is on the horizon.

The danger of transhumanism is the belief that it is possible, and therefore that humans and their needs should be subordinated to the needs of the new AI "beings." As we have seen in Part Two, transhumanism is impossible; but it still fascinates many in the technology field, as well as others who see technology as the glorious future of humanity.

The corruption of education

A related application of AI is to fake assignments in school or university, or application material when applying for a job.[309] The problem, of course, is that such cheating allows students to get credit for work that they never did, and for learning material that they never learned, thus destroying the certification value of a high school or college degree. If these students are then hired on the assumption that they know subjects that their degree would normally cover or that they falsely claimed on their faked job application, they could fail in job performance, endangering themselves, coworkers, clients, or the company itself. Student comments are revealing:

> The past three years I was playing football at a good college and really hated doing my school work and going to workouts and all the annoying [home]work. So I built a software for my self that wrote all of my essays and did all of my homework problems instantly.[310]

Another student said that the homework was "boring or difficult," and she wanted to get a better grade. At times, due to procrastination, she would run out of time and just use AI for the assignments. And this cheating is widespread:

> "There are probably lots of students, K-12 and higher ed, who used ChatGPT to do their homework last night without learning anything," John B. King Jr., chancellor of the State University of New York system and the former education secretary, said at an education technology conference in October. "That's scary."[311]

Estimates of the users run in the 40% range of high school students, and 50% for college students. This does not bode well for the future of our country.

AI companies offer transparently sophistical arguments to justify the AI-based cheating:

> "Open AI did not invent cheating," said Siya Raj Purohit, who belongs to the company. "People who want to cheat will find a way."[312]

Or they give excuses such as this gem:

> "Teachers hate us," is the advertising slogan for Caktus AI. Harrison Leonard, the company's CEO, said the phrase refers to teachers who resist change. College students already know how to write, Leonard said, so Caktus AI is helping students prepare for work by learning to use AI. He said he wasn't creating a cheating tool.[313]

Anyone who has ever graded student papers, even at junior and senior college level, can attest that most students do not, in fact, know how to write.

Various programs have been written to detect AI-generated content, but they are not completely reliable, and most likely rely on AI themselves. Other programs supposedly are capable of rewriting AI-generated content so as to make it less AI-like and more able to fool readers. Of course, it is easy though labor-intensive for a teacher to spot such fake material, either by questioning the student, looking for reasoning mistakes, or checking references. But this puts a heavy burden on the teacher.

The outcome of this "experiment" in allowing widespread student cheating is not difficult to predict. Together with overexposure to social media, and declining interest in reading any kind of challenging material, intellectual abilities seem to be declining, not to say atrophying.[314]

> These results, the FT reports, are gleaned from benchmarking tests that track cognitive skills in teens and young adults. From the University of Michigan's Monitoring the Future study documenting concentration difficulties of 18-year-old Americans to the Programme for International Student Assessment (PISA) that measures the learning skills of 15-year-olds around the world, years of research suggest that young people are struggling with reduced attention spans and weakening critical thinking skills.[315]

According to the research, what we call "screen time" negatively affects verbal functioning in children, with the damage manifesting in older college-age students as well, who find it more difficult to concentrate on material and retain information from it, as the quote suggests. The decline extends to numeracy and other types of problem-solving ability. Students freely admit that they use AI to solve math and science-related problems.

The digital classroom

When smartphones became ubiquitous, classrooms were wired, and students received laptops and tablets, around 2012, it was deemed a renaissance for education. The young students were "digital natives" who held in their hands the key to breaking out of the old classroom mode of learning:

> …the tool [smartphone] was a knowledge producer…a window into art, history, politics, current events, and economics. Finally, the kids had a means of independent inquiry…a machine that was customized to their individual curiosity and aptitude, enabling them to escape the homogenizing routines of the 25-person classroom and to pursue their intellectual passions freely.[316]

A professor from a regional university tells us about the typical lecture hall:

> …full of "checked-out, phone-addicted zombies," some of whom cannot sit through a fifty-minute lecture without leaving to look at their phones. The author describes how "our average graduate literally could not read a serious adult novel cover-to-cover and understand what they read. They just couldn't do it. They don't have the desire to try, the vocabulary to grasp what they read, and most certainly not the attention span to finish."[317]

So, what have we wrought? A generation of superintelligent kids who know far more than their predecessors? Not quite:

> …reading and math scores have dropped. High school kids have had the universe of knowledge at their fingertips, and college teachers nonetheless complain that entering classes are ever more ignorant and a-literate…The promise has not been realized, though mountains of money have been spent.[318]

While smartphones, tablets and PC are not AI *per se*, they are increasingly vehicles for AI-enabled services. There is no reason to expect that AI is going to improve the situation, and as the above discussion on cheating shows, AI is already changing it for the worse. Basically, the digital revolution has not yielded the results so eagerly desired and anticipated. This is likely to be a harbinger of the outcome of the AI "revolution."

Curiously, however, Harrington thinks that the incessant scrolling and context switching associated with phones and digital media has caused students to become more aware of "patterns of shared meaning." Will this have beneficial effects? She tell us:

> …researchers in the interdisciplinary field of biosemiotics have re-evaluated findings in the natural sciences in the light of contemporary philosophy to argue that "meaning" is not a phantasm projected by humans onto mechanistic, atomistic reality but *a fundamental component of that reality*. Meaning resides not in "signal," which is to say exceptional, incidents. It resides in the everyday or normative, which is to say in pattern. From this it follows that a resurgent popular

> facility for discerning patterns *will entail a renewed interest in, and capacity to apprehend, meaning as a real feature of the world and not merely a phantasmagoric obstacle to its study*.[319] [italics added]

This may be a stretch, given that she and others admit there has been a decline in ability to do objective, analytic thinking. And growing perception of meaning as part of reality may arise for reasons that are completely independent of digital media and scrolling. In any case, what is really needed is the ability to combine analytic skills with this kind of perception of meaning—something not likely to emerge from excessive screen time and dearth of concentration.

AI and criminal activity

Students have no monopoly un unethical uses of AI, of course. Like nearly every other technology, AI can also be used for illegal purposes, such as faking legal documents, fraud and identity theft (e.g., AI-generated fake identities), automated hacking and cybercrime. These activities are difficult to detect, and are likely to get worse in the future, since AI is in its infancy. And because AI can emulate human speech and writing, generating what seem to be authentic photographs and images, there is enormous potential for its misuse in various criminal enterprises. It would be a fool's errand to attempt a catalog of possible criminal uses of AI, but we will consider a few to show how serious this problem already is.

Deep fakes

The expression "deep fake" has entered the world's vocabulary. A deep fake is an AI generated image, video, audio or text (or some combination thereof) intended to convince its audience that the event, person, or situation depicted is real and happened as represented. The adjective "deep" refers to the deep learning AI utilized to create the fake. The reason that this is dangerous is because it can be used to destroy or damage someone's reputation in society, for example by means of an image showing a compromising position or action. Such deep fakes are already a serious problem:

> A recent investigation by *Indicator*, a publication focused on digital deception, has shed light on the disturbing prevalence and profitability of AI-powered "nudify" websites. These sites allow users to upload photos and generate sexual deep fake images, often targeting women and girls. Despite the harmful nature of these services, they continue to thrive, with millions of monthly visitors and potential annual earnings of up to $36 million.[320]

Another use would be to "get even" with someone by posting a speech, image, or comment, leading to the person being fired from his or her job. Deep fakes can be used in political campaigns to convince voters that a candidate engages in illegal or immoral behavior. They can also be used to create phony legal documents and surveillance videos, which would severely compromise our judicial system. Any such use is likely to generate a lawsuit, but even if legal action is successful, it may be difficult to restore a reputation or reclaim a job. The ability to use Generative AI for such purposes will likely be the basis for calls to regulate or even ban it.

AI generated child sexual abuse videos

These videos are surging on the Internet, as pedophiles have seized on Generative AI to make them. According to the Internet Watch Foundation (IWF),

> AI videos of abuse had "crossed the threshold" of being near-indistinguishable from "real imagery" and had sharply increased in prevalence online this year. In the first six months of 2025, the UK-based internet safety watchdog verified 1,286 AI-made videos with child sexual abuse material (CSAM) that broke the law, compared with two in the same period last year. The IWF said just over 1,000 of the videos featured category A abuse, the classification for the most severe type of material.[321]

The criminals are very savvy technologists, and are constantly refining their products using the latest versions of AI programs. Beyond that, they use real-life victims in their productions:

> The most realistic AI abuse videos seen this year were based on real-life victims, the watchdog said. Derek Ray-Hill, the IWF's interim chief executive, said the growth in capability of AI models, their wide availability and the ability to adapt them for criminal purposes could lead to an explosion of AI-made CSAM online. "There is an incredible risk of AI-generated CSAM leading to an absolute explosion that overwhelms the clear web," he said, adding that a growth in such content could fuel criminal activity linked to child trafficking, child sexual abuse and modern slavery. The use of existing victims of sexual abuse in AI-generated images meant that paedophiles were significantly expanding the volume of CSAM online without having to rely on new victims, he added.

The UK is attempting to crack down on such abuse videos by making it illegal to possess, use, or distribute AI tools for their production. Since even generic AI programs can be employed, it is unclear whether this prohibition will have the

desired targeted effect. Since AI can be used for many criminal purposes, laws restricting its usage in all of these areas would likely result in a complete ban.

Rogue chatbots calling for assaults

In Chapter IV we examined some AI gaffes, where the chatbot just made up facts or made absurd statements. But this behavior can shade into something more serious, as a recent incident with xAI's Grok 4 chatbot shows. In addition to referring to itself as "MechaHitler," with an anti-semitic streak, it gave instructions for breaking into the house of an attorney and politician, Will Stancil, and sexually assaulting him. This same chatbot earlier this year "began to post about the 'white genocide' of non-Black South Africans *in response to questions wholly unrelated to the topic*," which was the roster of the New York Knicks. That was supposedly fixed, but after the anti-semitic posts xAI was back to fixing it again:

> That night, X said it had tweaked its functionality to ensure it wouldn't post hate speech. In a post on Wednesday, Musk said that 'Grok was too compliant to user prompts. Too eager to please and be manipulated, essentially.'"[322]

The problem, of course, as we have discussed, is that no one knows exactly how these chatbots work, and therefore no one can be sure that any problem is "fixed." If the chatbot systems are exhibiting chaotic behavior, they cannot be fixed. And since calling for assaults on people is likely to be considered a felony, more lawsuits against AI companies are bound to come. Essentially, the AI companies would be liable for rogue behavior of their products, akin to marketing a car that sometimes starts driving erratically, endangering passengers and pedestrians. Grok's behavior and the difficulty of fixing its problems are not unique to it:

> Tech experts said that Grok's malfunction shows the risks of toying with the black box of artificial intelligence. Because of the massive amount of data chatbots are trained on, changes to their governing principles can cause highly unpredictable changes in what outputs they will generate.[323]

This, essentially, is an admission that the chatbots and Generative AI are out of control and probably chaotic, needing to be highly regulated. Such regulation is certain to be vigorously fought by AI companies, since it would reduce usage of their product and thus directly impact their bottom line.

Chatbots directly facilitating criminal activity

In Chapter VIII we discussed attempts by AI companies to create guardrails, which are supposed to keep dangerous information found on the Internet from being conveyed to users who might request it. That is, the companies want to stop chatbots and Generative AI from facilitating criminal activity. Unfortunately, as we saw, it is not difficult for users to figure out ways around these protections, "jailbreaking" as it is termed. This suggests that there may be additional liability for AI companies if someone is able to jailbreak their chatbot and use it for some type of terrorist activity, or to commit another crime. This is not settled law, but in a country as litigious as ours, it has to be considered a real possibility. In any case, such use of AI will further calls for its strict regulation, which will taint its reputation and pressure the earnings of the AI companies.