Published in *Artificial Intelligence for Better or Worse* in 2019 by Future Leaders (www.futureleaders.com.au)

Complexity is not new: how our own technological history can teach us about AI

Elizabeth T. Williams, Caitlin M. Bentley, Katherine A. Daniell, Noel Derwort, Kobi Leins, and Ehsan Nabavi

> *A certain ruthlessness is encouraged by the mistaken belief that to disregard human considerations is as necessary in technology as it is in science. The analogy is false.*
>
> *Hyman G Rickover*[1]

What do the words "artificial intelligence" evoke for you? Hopes? Fears? A shiny, personalised future with a place for everyone? A dystopian landscape, peppered with fallen drones and unemployed masses? Or perhaps the term evokes nothing more than the world we already live in.

Artificial intelligence (AI) is already pervasive; it helps shape cities, homes, and increasingly, lives. It is a term often used to describe a vast collection of technological systems sharing certain properties. These systems have sensors, or a means of collecting data about the outside world. They have actuators — again, broadly defined here as some means of enacting change in the outside world. They have a collection of computational objects (software and hardware) that enable the system to take the sensor inputs, interpret them, make decisions based on them, and bring the actuators to life. These computational

objects have the capacity to learn, meaning they can adjust their behaviour in response to the environment in which they find themselves — at least, based on the sensor data they are capable of collecting.

It is this last property — the capacity to learn — that captures our imagination. If a system can learn, it does not have to be preprogrammed for every scenario it may encounter. We can create more efficient code, design more flexible systems, tailor systems to a particular individual or situation, or send systems out to navigate unknown territories. The capacity to learn creates possibilities.

But with those possibilities comes a loss of control. How do we make sure these systems do what we want them to do? With this, there is a directly related question: how do we keep them from doing things we don't want them to do?

The AI we have today is 'narrow' — it is created to do a particular task according to specifications its creators define. While it can optimise its ability to perform a given task, it cannot evolve to do another task. There can be consequences – sometimes serious — when such a system is poorly designed, but there is still a sense of control. There are ways to set a variety of guardrails on the system and optimise those guardrails until they work. Journalists can investigate and inform the public, people can call for reform, governments can regulate, and creators can change or decommission the errant AI system.

The AI we might imagine when we worry about control is general – (super)human-like in abilities, if not in outlook. If the system is smarter than humans, one might presume it will find a way around any guardrails placed in its path in its quest to

pursue its goal. This type of AI it is currently either a long-term research question or a myth, depending on who you ask, but there's a lot of time and development in between the 'narrow AI' of now and a (possibly imagined) future in which this 'generalised AI' exists.

How will the AI we create now scale to different places, times, and uses – and how much influence will AI have in developing its own future? How will human society maintain control over this technology as it continues to evolve?

In some respects, this question of control is not new. We have been building systems of increasing complexity — and risk — throughout human history, sometimes learning (and sometimes not) from our missteps along the way. In this chapter, an accident involving one of these complex systems — Three Mile Island — will help to illustrate where and how creators of systems involving AI might begin to seek insight.

Three Mile Island: The story

We will turn now to the scene of the accident — a flat slip of land surrounded by the Susquehanna River, near Goldsboro, Pennsylvania, in the United States. It's just before 4 am on 28 March 1979. Darkness has long settled over the small towns on either bank and the birds have not yet woken from their slumber. Night shift workers at the Three Mile Island nuclear power plant are going about their duties as usual — clearing pipes, monitoring temperatures and pressures, checking off tasks. Suddenly, in the control room for one of the two reactors on site, TMI-2, alarms begin to sound.

Operators in the TMI-2 control room scramble to assess the problem in the midst of a cacophony of noise they cannot

silence. In the secondary coolant system — a series of water-filled pipes, pumps, other equipment that help keep the reactor sufficiently cool and use the fission-generated heat to make electricity — the main feedwater pump has stopped working.

A backup to the secondary cooling system should have kicked in — its pumps are working, after all. But the pressure and temperature in the reactor core continue to rise. In the chaos, operators miss two control room lights — one obscured by a yellow maintenance tag — that indicate two valves isolating this backup from the secondary system have been manually closed. The backup is therefore quietly useless, but for now, no one notices.

Problem 1: The valves between the secondary cooling system and its backup are shut, rendering the backup system useless

A valve on the primary coolant system — the water system in contact with the fuel rods — opens automatically, to decrease the pressure of the system. This pilot-operated relief valve, as it is known, should shut again to keep the primary system sealed, but the valve mechanism malfunctions, leaving the primary system open. In the control room, a signal leads the operators to believe the valve opened, then shut, as expected.

Problem 2: The pilot-operated relief valve remains stuck open

By now, the reactor has scrammed, which means the control rods have dropped in automatically and the fission chain reactions produced within the reactor core have halted. This slows the temperature rise and buys the operators time to figure out how to regain control of the pressure and temperature in the primary system — both of which are crucial for protecting

the integrity of the fuel rods within the reactor core. But there is still no new cool water flowing in, and the radioactive by-products of fission will continue producing excess heat for some time. With the pressure still rising, another emergency system kicks in — this time, a set of high-pressure injection (HPI) pumps responsible for topping up the primary system. The HPI pumps begin pushing a steady stream of cool water into the primary system.

Soon, though, an operator notices something else to worry about. The pressuriser — a tank in the primary system that sits above the reactor core, which usually has a mix of liquid water and steam and is used to control the pressure in the primary system — appears to be filling up with liquid. Operators presume this is a result of the water the HPI pumps are introducing to the primary system. This is a cause for concern because this system is hard to control when it is full of liquid water, so the operators decide to turn off one of the HPI pumps and slow the flow of the second one down to almost nothing. This is a mistake.

Problem 3: The operators alter the primary backup HPI pump flow rate and function

The sensor providing the operators with information on the pressuriser water level is functioning properly, but it cannot detect the difference between a scenario in which the water level is high in the primary system and one in which liquid water is being pushed up and out of the primary system by steam. The second scenario is reality, and is distinguishable from the first, but only if the operators think to also look at the reactor pressure and temperature together and as a function of

time. Only these observations together with the pressuriser water level can reveal the state of the water in the primary system.

Problem 4: The operators fail to use the pressure and temperature readings as a function of time to understand the current state of the water in the primary system

Soon, the primary system begins boiling dry through the open pilot-operated relief valve, leading to what is known as a "loss-of-coolant event". The fuel rods within the reactor core begin to rupture. TMI-2 will never produce electricity for the US Eastern electrical grid again.

Normal accidents, complex systems

> *We are convinced that if the only problems were equipment problems, this Presidential Commission would never have been created. The equipment was sufficiently good that, except for human failures, the major accident at Three Mile Island would have been a minor incident. But, wherever we looked, we found problems with the human beings who operate the plant, with the management that runs the key organization, and with the agency that is charged with assuring the safety of nuclear power plants.*[2]
>
> *President's Commission on the Accident at Three Mile Island*

The accident at Three Mile Island is what Professor Charles Perrow calls a "normal accident" — an accident in a complex system with many interlocking components working over short

timescales. His investigation of these phenomena began with the accident at Three Mile Island.[3] These accidents often begin with some small, seemingly insignificant problem, the consequences of which quickly spiral out of control because of the complexity and interconnectivity of the system in which the problem occurs.

The accident may have been avoided if any one of the problems identified in our story had been rectified. If the operators had only realised the valves between the secondary system and its backup were shut, they could have opened them. If they had realised the pilot-operated relief valve was open or simply kept the HPI pumps working as designed, they could have avoided the loss-of-coolant incident that triggered the fuel rod damage.

If the operators had realised …

Do you notice any semblance of blame creeping into your thoughts now? Hindsight is tricky this way. Our first response is to blame the humans closest to the accident — the ones that "should have known". But should they, in fact, have known?

We have focused on the first few minutes of the Three Mile Island accident here for brevity's sake, but we could have drawn back the curtain in time and space to reveal increasing complexity — not just within the reactor on site that day, but in all the technical, social and political systems that influenced and were in turn influenced by that system and its evolution. The operators were part of this complexity and were a product of it, in many respects. Their actions were based on experience gained during normal operations and training for how to handle major disasters, based on guidelines produced by the plant designers, regulators, and management. Their actions

were also shaped by actions taken (or not taken) by the automated safety responses of the engineered system they were charged with operating.

The string of events that day was not normal, nor was it a major disaster — until all the minor incidents together added up to a major accident.

Human-machine partnerships

Let's connect this classic example of a normal accident back to our current exploration of AI. Consider an AI system with the potential to take actions that may be seen as risky. Those involved in shaping this system tend to prefer that humans have the capacity to intervene in the actions the system takes. This choice is made because no one can be sure they have accounted properly for improbable high-risk scenarios a system might one day encounter in the world.

Humans are our solution, however imperfect, because we still have faith that we mostly understand how humans make decisions. They can respond to novel situations with proper training and explain their actions afterwards. To return to the example of Three Mile Island, we can imagine taking on the role of those involved in reviewing the accident and hearing the operators' stories of what they did and why. We can put ourselves in their place and learn from their mistakes, with the idea that we can prevent future accidents.

This is in interesting contrast to a typical engineering approach that prefers — wherever possible — to build automatically triggered safety systems that don't require human action to work — and may not (as the service tags at TMI-2 reveal) adequately take into account possible human interventions.

Humans aren't "trusted", but they are — mostly because there is currently no satisfactory alternative choice. As a result, a partnership between human and machine typically is forged.

The way system creators, managers, and regulators think about and plan for effective human-machine partnerships in any complex system is relevant to the questions we began with: How do we make sure a system does what we want it to do? How do we make sure we keep them from doing things we don't want it to do?

Sparse data, imagined realities

There is another parallel to draw between AI and Three Mile Island. An investigation of the accident at TMI-2 revealed that the humans in the system were acting as well as they could have in the face of partial information. The operators did not have full knowledge of the state of the reactor that day. They could not have — there was no window into the environment within the reactor core, no direct sensor information that would have provided a realistic assessment of what the system was doing. They were not unlike an AI[4] trained on a dataset filled only with data representing either normal or catastrophic conditions — not the small chain of problems faced by the TMI-2 system on the day of the accident.

There are also possible comparisons we could make to an AI system that cannot take sensor data capable of illuminating an emergent problem.

The possible realities suggested by the data the system collects have an influence on actions the system chooses to take. This is shaped, in turn, by the possible realities system

creators (and influencers) imagined when designing the system in the first place.

Whose imaginations and experiences are embedded within the design of the system? When and how might that collection of imagination and experience fall short of reality, and what are the possible consequences?

Defining a goal

> *The most serious "mindset" is the preoccupation of everyone with the safety of equipment, resulting in the down-playing of the importance of the human element in nuclear power generation. We are tempted to say that while an enormous effort was expended to assure that safety-related equipment functioned as well as possible, and that there was backup equipment in depth, what the NRC [US Nuclear Regulatory Commission] and the industry have failed to recognize sufficiently is that the human beings who manage and operate the plants constitute an important safety system.*[5]
>
> *US President's Commission on the Accident at Three Mile Island*

Let's step back from Three Mile Island for now and return to one of the questions we began with — but this time, let's make our queries more specific. How do those with the power to shape an AI system begin to make sure it does what they want it to do — both now and as the system scales to different settings, times, and societies? And how do they keep the system from doing things they don't want it to do, even as it extends its reach

beyond the places and times for which it was originally designed?

To explore this question, let's start by stepping back and considering the starting point for many AI systems in existence today. To craft and shape a system, the system creator(s) typically have had a goal in mind. What do you want your system to achieve? What vision do you have for it? What change do you see it affecting in the world, and what shapes your perception of the world you want your system to impact? There is usually a why in the creator's mind — a reason to make the effort — but let us set that aside for the moment and focus on the goal itself.

What influence might the goal have on the system as it evolves? What influence might the system's evolution have on the goal? And which of these questions has primacy? How have the goal and the system changed with time, setting and place? Is the system that has resulted from that evolutionary process consistent with the goals the system is reportedly trying to achieve — and if not, how has that come to be?

There are feedback loops here, loosely described as mechanisms of mutual influence between a system's design, actions, goals, and context(s). To understand these feedback loops, it helps to gain a sense of where a technology came from and how and why it has taken its place in our lives. There is usually not one story here, but many.

For Three Mile Island, the goal of the system was always to generate energy safely, but the context of the technology at the heart of the story and the actors involved in shaping that technology for human use had shifted dramatically over the

course of its existence. One could tell the story of the evolution of the pressurised water reactor, first designed for use in US Navy submarines. One could ask questions about the changing influence of key actors — for example, Hyman George Rickover, the man often credited with leading the US Navy and US civilian commercial world into the business of nuclear energy, who expected any of his personnel working on nuclear reactor technologies to uphold high standards and have a deep understanding of the systems with which they worked.[6] One could also tell the story of a newly formed US Nuclear Regulatory Commission charged with ensuring the safety of nuclear technology, full of personnel still accustomed to helping the relatively new nuclear energy industry thrive. One could ask questions about why humans were not explicitly considered part of the reactor's safety system — and why their training did not prepare them to do their job adequately.[7]

For AI, these stories may be indirectly or directly embedded within the technology via the human-provided (and categorised) data that shapes the models, the human decisions that shape the decision-making code, and the implicit assumptions built within all of these things that trace back to the origin stories of a particular system. These stories matter in assessing the question of control. Why was the system developed? What were (or are) the intentions of its creators? What are the intentions of its users? What assumptions did we built into it? Do these assumptions still apply, and if not, have we adjusted the system or its use appropriately? How does the system (and this includes any relevant human actors) work to achieve its goals? Who or what is involved in taking action to achieve a system goal? What possible actions could they take? What information

and past experience or data might shape their actions? And what are the possible consequences of all those actions, over what timescales? What are their potential blind spots? What guardrails have been placed within and around it, and how do they work?

In exploring these questions, we can begin to develop a deep understanding of the system and its imagined place within our world. We can understand why it was made in the first place, and imagine how it might achieve its goals, and what the consequences of that pursuit in the face of changing context might mean.

But there's another piece we haven't considered fully: the rest of the world.

The environment, the people, the other systems (complex or otherwise) that can and will interact and possibly intermingle with our imagined system over the course of its lifetime are all a part of this world. What about all of these other actors — also possibly pursuing goals, also independent, also capable of having their own intentions and perceptions about when and how they might interact with the rest of the world and how the rest of the world might interact with them? How do they influence our system? How does our system influence them? When and how might this matter?

All of this potentially adds up to an incredibly complex problem. How do we transform all of this information into a meaningful path forward?

Towards a new applied science

The hypothetical AI system we are imagining here exists within a web of complex social, technological, and environmental

systems, but even this is nothing new. There are many known ways of understanding complex systems, and many ways of using this information to inform how such systems are shaped.

At the heart of these methods lies a series of questions: What is the system? What do we include within it? What are its boundaries? Where does the system begin — and where does it end? What should we include, and what can we comfortably ignore?

With time and experience, one can only come to a single vexing conclusion to all these questions: it depends. The system and its boundaries shift based on the questions we are asking, and on what we intend to do with the results of our inquiry.

All is not lost, though. There are methods for this, too. What seems missing now is a means of adapting and applying this knowledge in an actionable way to AI — and helping AI adapt in turn.

So where do we begin?

Here at the 3A Institute, we've been exploring this question. Our aim is to create a new applied science that will help AI scale safely and equitably. We began with a team of researchers with incredibly varied backgrounds — from sociology and anthropology to computer science and nuclear physics — who were willing to work together towards this single mission. We started exploring well-documented case studies like the Three Mile Island accident, to see what we could learn. We brought on industry and government partners willing to give us insight into new or emergent AI systems they were considering or using. We took on a cohort of Masters students with similarly diverse backgrounds and outlooks, who were willing to spend a year

helping us define what this new, as yet unnamed applied science would look like.

Because of all of this work together, the new applied science we envisioned is now starting to take shape.

For better or worse?

So, AI: for better or for worse? Our relationship with all of the technologies that could be called AI is complex, varied and everchanging. AI is difficult to comprehend because it can be everything from a highly sophisticated energy distribution grid to a small, self-contained robotic toy bug with just enough intelligence to allow it to navigate across the ever-evolving landscape of your living room. It can be built on snippets (data) from the past, and by people from very different places, cultures, and assumptions. It usually involves more human input, and certainly more human training, than you may at first imagine. Today, it can only perform a specific task; in the future, it may provide us with a technological counterpart of equal or superior intelligence.

Right now, Rickover's words at Georgetown University's Symposium on Cybernetics and Society back in 1964 still hold truth: "We alone must decide how technology is to be used and we alone are responsible for the consequences."[8] It is humanity's job to understand the AI systems we create or adapt, and to think carefully about how we use them. It is also our job to question, manage, and find ways to appropriately govern the systems we create to help AI scale safely and equitably now and into the future. All of this depends on having a more nuanced discussion of what AI is, where humans are truly involved in its

design, use, and function, and how the complex systems that preceded it have influenced its development.

Endnotes

1 Rickover HG (1965). A humanistic technology. *The American Behavioral Scientist,* January, 3–8.
2 Kemeny JG et al. (1979). *Report of the President's Commission on the Accident at Three Mile Island: The Need for Change: The Legacy of TMI.* Lessons Learned Information Sharing (LLIS).
3 Recommended further reading: Perrow C (1999). *Normal Accidents: Living with High-Risk Technologies*. Princeton University Press.
4 We have machine learning in mind for this example.
5 Kemeny JG et al., op. cit.
6 Duncan F (1990). *Rickover and the Nuclear Navy: The Discipline of Technology.* United States Naval Institute.
7 Kemeny JG et al., op. cit.
8 Rickover HG, op. cit.