Artificial Intelligence in Critical Infrastructure Systems

Phil Laplante, Penn State

Ben Amaba, IBM

Seven expert panelists discuss the use of artificial intelligence in critical infrastructure systems and how it can be used and misused. They also address issues of public confidence in such systems and many more important questions.

The U.S. Cybersecurity Infrastructure and Security Agency (CISA) defines 16 critical infrastructure sectors: chemical, commercial facilities, communications, critical manufacturing, dams, defense industrial base, emergency services, energy, financial services, food and agriculture, government facilities, health care and public health, information technology, nuclear reactors, transportation systems, and water and wastewater systems.[1] Together, these sectors make up the set of "critical systems."

But what happens when artificial intelligence (AI) and machine learning (ML) are incorporated in such systems? How do we ensure the safety of the public and critical infrastructure? How can we ensure the public that these systems will be safe?

In this virtual roundtable, we asked seven experts across many of the critical systems domains to address these and other questions. (See "Roundtable Panelists" for more information about the panel.) Their answers are diverse, comprehensive, and sometimes surprising. Answers are given starting in last-name alphabetical order and then rotated, circularly, through the list.

COMPUTER: What does "AI in critical systems" mean?

JONATHON BARKLEY: The operation of a critical system normally requires a human intervention to detect/diagnose a problem and then make a decision and/or take action, for example, acting as a pilot or a surgeon. But even a trained human can be inefficient or make mistakes due to various psychological conditions; this is where AI can play an important role, eliminate such mistakes, and be more efficient than humans.

Digital Object Identifier 10.1109/MC.2021.3055892
Date of current version: 24 September 2021

ROUNDTABLE PANELISTS

Jonathon Barkley is the R&D head and group (global) director at Moog Medical Devices. Contact him at jonathanbarkley@gmail.com.

Jeff Daniels is the director of engineering, manufacturing, and automation at Lockheed Martin. Contact him at jeff.daniels@lmco.com.

Cliff DeBerry is the vice president of design, construction, and delivery at Memphis Light Gas and Water. Contact him at cdeberry@mlgw.org.

Bart Kemper is a principal engineer at Kemper Engineering Services LLC. Contact him at bkemper@kempereng.com.

Andrei Popa is the reservoir management digital advisor with the digital platform team at Chevron Technical Center. Contact him at andrei_sp@hotmail.com.

Matheus Scuta is a product manager, operations analytics, at Ford Motor Company. Contact him at matheusscuta@gmail.com.

Kent Welter is the chief engineer, testing and analysis, for NuScale Power. Contact him at kwelter@nuscalepower.com.

JEFF DANIELS: We can refer to the CISA 16 critical infrastructure sectors whose assets, systems, and networks are considered so vital to the United States. These systems are increasingly software defined; software enabled; and connected through sensors, beacons, and multiple communications (5G, WiFi-6, Bluetooth Low Energy, radio-frequency identification, and so on).

As we collect telemetry data and apply AI to help operate and manage critical systems, we need to be conscious of the ethical practice of AI. The disruption, incapacitation, or destruction of critical systems infrastructure could have adverse effects. The challenge is managing the bias in AI models and effectiveness.

One example is that an "idle" machine may appear to not be in use by only looking at the data; however, the operator may actually be setting up a job to execute on the machine. We need to be very clear on the operational modes, data lineage, and applicability to critical systems operations.

CLIFF DEBERRY: AI means Memphis Light Gas and Water can leverage this emerging technology to continue to offer safe, cost-effective, and reliable services to the customers we are privileged to serve. While our efforts have concentrated on data-driven applications that are customer facing, the ability to leverage the technology as a predictive tool to forecast potential failures that would cause outages or increase cost provides an opportunity for exploration.

BART KEMPER: "AI" is a numerical modeling system that is considered to approximate human judgment or rational thinking (two different things), although, in current applications, it's understood that the AI "expertise" is in a limited scope. "Critical systems," in this context, are subsystems or components of a larger system such that the loss of function of a "critical system" results in a loss of function of the system as a whole. To answer the question, this means that an AI was used to replace or supplement human judgment and actions, be it an individual or a group of humans, in a system such that a failure of the AI (or humans) results in a failure of the system to function.

An example of this is humans performing air traffic control at a major airport. The humans are continually updated regarding the air and ground conditions, and they use that to make assessments of how to give guidance to aircraft in their airspace while giving consideration to schedules (commercial significance), safety (paramount criteria), and other factors. Humans can make errors of a variety of types, which, in turn, requires checks and balances within the system to keep an individual from being a single point of failure to properly give guidance to an aircraft.

An AI system could update its data set more quickly, more often, and in more detail as well as handle a greater number of aircraft than a single person, opening the door to reducing the human workforce. However, once these people are replaced, a failure on the AI's part is less likely to be detected due to less oversight and will be of higher significance than a single person, as it has replaced multiple people, which, in turn, means the system will not be able to maintain the same capacity without the AI since the replacement staff is not in place.

ANDREI POPA: Critical infrastructure systems are those systems that directly affect the public's health, safety, and welfare and whose failure could cause catastrophic loss of life, assets, or privacy. Examples would include energy, distribution, communications, health care, and financial services. These systems have become increasingly dependent on AI technologies to automate many of the routine tasks that humans used to perform—however, at a speed and objectively well above human capabilities.

MATHEUS SCUTA: A critical system is a system that requires extremely high reliability. Some critical systems, such as pipelines and electrical grids, are generally overlooked/taken for granted by the general population until they must pay attention to them. AI can play a huge role in maintaining the reliability, evolution, and performance of the critical system. AI in a critical system is a necessary advantage to ensure the success of the system's performance by enabling the critical system to adapt to any variations without disruptions to the end customer.

***COMPUTER*:** What does AI in a critical system look like to a user?

DANIELS: AI within critical systems should be seamless, fully integrated, and embodied in the "flow of work." We must include the ability for ML and neural networks to capture heuristics from "human-in-the-loop" operations and adjust accordingly.

If AI is treated as an afterthought, sidecar, or bolt-on application, it will not scale or be readily adopted. We must also consider the feedback loop for operations in terms of the timing and process in split-second decision making, specifically where the human handoff occurs. AI must be included in the systems, processes, and operations in how we work, that is, a seamless user experience.

DEBERRY: Users of the technology will see a tool that easily integrates into their everyday routine. From better utilization of resources, prediction of peak demands, and forecast of potential power failures to lower utility bills and customizable services, users should experience cost reductions in the delivery and consumption of energy.

KEMPER: In broad strokes, it will be invisible or a "coworker." It's invisible if the user is downstream of the AI's activity and being given the results as something to act upon. It would be no different than getting the output of a team of people. If the user is at the same operational point as the AI, like a coworker, then he or she might get a screen full of data with an "AI recommendation" given, highlighting that this is the result of the AI but still having human intervention at that point in the process.

POPA: To help you understand what AI in a critical system may look like to the end user, I would like to use an example from the energy sector—the U.S. infrastructure pipeline network. The advances in AI led to not only a process automation but also the development of a series of advanced and real-time monitoring of pipeline network systems that significantly improved the reliability, operability, and cost optimization of the fluid transport and delivery.

One area is the use of drones that fly over the pipelines, streaming image data to a base system where AI deep learning models can perform real-time pattern-recognition tasks and identify abnormal conditions, such as damage, corrosion, leaks, and so on. Now, consider hundreds or thousands of drones that can stream the data while the AI performs this task objectively, 24 h a day, without any interruption. While the AI is completely invisible to the end user, the system generates an alarm when it detects abnormal operating conditions that allows the user to act.

Furthermore, some systems are embedded or linked to shut-in safety procedures that can close or open a valve or bypass to avoid potential incidents. The end user in this case would only monitor the execution procedures.

SCUTA: It depends; it can be completely "behind the curtain," having no actual apparent impact to the end user or end consumer, or it can also be a major component of the customer-facing solution. My favorite "behind the curtain" example is how a power grid can respond to an increase/decrease of electricity consumption by increasing or decreasing the electricity availability according to shifts in demand.

KENT WELTER: AI in a critical system is often transparent, meaning that the expert system is typically designed to provide automated or automatic actions with minimal feedback to the user.

BARKLEY: Here are three examples:

- *an autonomous vehicle*: a self-driving car, train, or aircraft
- *a diabetes management app*: a technology that can track all of the activities and guide a lifestyle dynamically
- *a nuclear plant monitoring system*: a system to constantly take various inputs from sensors and take actions to avoid disasters.

***COMPUTER*:** What are some of the enabling technologies?

DEBERRY: Virtual agents, such as chat bots, allow FAQs to be addressed without the need of a live agent as weak as natural language generation that converts speech to text is. Biometrics is valuable in granting access to secured locations/entrances.

KEMPER: In addition to neural networks and ML, a key enabler is sensors. A conclusion is influenced on the depth and breadth of the supporting data. The ability to put more sensors for a given purpose to measure more things at greater resolution at a faster rate is a key enabler in maximizing the computational nature of AI over the "thin-slicing" learned intuition of a human (see *Blink* by Malcom Gladwell[3]). Another key enabling technology is the communications network. It's not just a question of what data are coming in from where; it's also a question of how much data and with what lag.

POPA: AI is being deployed on wide-ranging systems, from data centers to edge devices. Artificial neural networks can be found in all engineering domains: energy, manufacturing, and agriculture as well as medical and health care. Fuzzy logic is applied to almost any process control and is a powerful decision-making tool that deals with uncertainty and ambiguous information. Genetic algorithms are ideal for optimization and scheduling tasks in the transportation, transmission, and energy sectors. And finally, case-based reasoning has been applied to manufacturing, energy, and even architecture and law to leverage the existing knowledge domain and lessons learned from past cases or instances.

The more advanced types of AI, such as deep learning neural nets, were originally developed for image and speech recognition and found their way into all critical infrastructure systems. An example is seismic interpretation via convolution neural nets, which is being used in the energy sector.

SCUTA: I think the technical side of this question can be better answered by others in this roundtable. However, other enablers that are much needed include the following:

- *Awareness*: It is necessary to generate awareness on how delicate the balance of a critical system is as well as on how users or institutions play a role in shifting the balance. For example, the fuel shortage during the Colonial Pipeline hacking was caused by the panicking of the general population, predominately because people had no knowledge of how delicate the critical system was.
- *Design thinking*: How can the system be enabled to operate within the new or existing environment? Change is imminent and necessary. Critical systems must be able to evolve at the required pace without any tradeoffs (for example, increasing the number of electrical grids while curtailing funding for the same grid's cybersecurity).

WELTER: ML using neural networks is often applied to decision-making algorithms for the system to respond to events in a consistent and timely manner. The drawback is that these enabling systems often have a hard time dealing with scenarios beyond their "training," which limits their application to critical systems.

BARKLEY: In simple terms, there are two kinds—one that relies a lot on the existing data and the other that relies more on intelligence and creates its own data sets.

DANIELS: We are seeing neural nets and ML applied in areas, such as prognostic health management and condition-based maintenance, where AI models have the ability to predict maintenance actions to avoid failures before they are required.

COMPUTER: What are some of the more exciting ways that AI in critical systems can be used?

KEMPER:

- *Infrastructure*: All transmission infrastructure (power; data; water; fuel; and air, ground, and water traffic) are the lifeblood of a modern centralized society. AI can produce efficiencies in all of these transmissions, which will, in turn, reduce wasted time, energy, and materials. It will also increase the ability to avert at least some disruptions as well as minimize the disruptions that do occur, such as those due to hurricanes or ice storms. We are already seeing the first infrastructure adaptations on roadways, where Maryland is setting aside a lane for autonomous vehicles.[4]
- *Research*: In the launch of the U.S. "Ocean Decade" earlier this year (sponsored by the National Academy of Sciences and part of the larger United Nations "Ocean Decade"), it was repeated that there are challenges in mapping the 3D space of the oceans and atmosphere. Mapping the landmasses, temperatures, currents, and chemical makeup over time is critical to understanding the climate. It was stated that current climate models are incomplete because they do not accurately track the heat transport associated with the water cycle, as water is the most influential of the "greenhouse gases."

  In several presentations, AI was discussed as being a key unfulfilled need by a number of researchers to give them data processing during the various research phases in order to use interim results to reassess and reassign sensors as well as process the data after the field phase. This is just one example of the research aspect with the physical sciences that would go toward immediate needs in addressing climate, food production, and pollution control.
- *Defense*: Technology is providing techniques like "swarm attacks," which are designed to defeat conventional methods by overwhelming targeting systems by using autonomous or semiautonomous guidance as well as numerical overmatch for the given target. AI is being looked at as a method to counter these attacks. It is also

being used to provide the detailed piloting for airframes, armored vehicles, vessels, and so on, with the human "pilot" merely providing general guidance, particularly with crewless systems.

In addition, AI is being used to detect and respond to various cyberattacks as well as in executing cyberattacks. Overall, the future of the military will be looking more like what was science fiction at the turn of the century, with crewless systems taking up more of the fight and AI being used for offense and defense in a multispectrum theater (which will include cyberattacks as well as AI-enabled messaging/information warfare).

POPA: One direction that will be exploited in the future is hybrid AI systems. These consist of the integration of multiple AI technologies that feed on each other to perform more complex tasks than one single technology can. As defined previously, the four AI technologies are exceptionally good for specific tasks. Therefore, a hybridization will lead to more advanced models with higher capabilities. An example within the oil and gas sector is reservoir exploitation through hydraulic fracturing, which addresses safety, the environment, and energy delivery.

AI could be used within this critical system to provide a real-time fracturing treatment design during job execution, achieving the desired target. If something does not go as planned during the pumping stage, an AI hybrid system redesigns the job in real time. A fuzzy logic controller for pressure matching will trigger a genetic algorithm optimization, which has a neural network model as the objective function. A complex task like this would bring AI as close as possible to human capabilities; however, its power of computation would be millions of times faster.

SCUTA: My favorite is the utilization of AI for the safety of the citizens of a country—such as facial recognition for police/immigration purposes. Another exciting capability is disaster response—knowing how to allocate resources (energy, emergency management services, and so on) to the best locations while monitoring the real-time effects of catastrophic events (human or nature made). For example, during an earthquake, what power grids or pipelines do we shut off to avoid a consequential fire?

WELTER: The application of AI techniques to predictive maintenance programs has the potential to save the nuclear industry millions if paired with big data.

BARKLEY: In the following kinds of applications: health care (surgical), high-speed transportation, financial, energy generation (nuclear), mining (underground), and military.

DANIELS: Emerging areas are in the generative design space, where the application of AI is designing products that no human would likely create. For example, when we look at the model-based systems engineering practice, we are starting to apply AI to build products based on certain characteristics. This approach is called *design for* X, where X could be any number of variables, including affordability, sustainability, cost, and so on.

DEBERRY: Data digitization that allows users to automate manual tasks, thus expediting decision making and predictive analysis so that we can be proactive in detecting potential failure rather than reactive.

***COMPUTER*:** What are some of the challenges to getting to these kinds of applications?

POPA: The development and application of an AI model relies on the ability to develop a high-performing model. Meeting this challenge requires a representative data set consisting not only of a large number of cases but also well-defined inputs that influence the outcome of the model. Furthermore, the other challenge we see today is the accuracy of the data being collected and used. Uncertainties in measurements as well as missing, biased, or skewed data lead to poor model performance with narrow application.

The other challenge that I see currently happening in the data science/AI space is the rush for the "shiny object." In the day-to-day environment, there is a lack of understanding of the problem to be solved.

I think that, before an AI technology or solution is proposed, engineers need to first understand the physical principles and fundamental laws of physics associated with the problem or task. Data science and AI are only enablers for problem solving and, when applied without a good framework and foundational understanding, lead to flawed solutions that bare no value and can, again, not only have consequences on critical systems but also erode the trust in the AI technologies.

SCUTA: I believe one of the biggest challenges is the security of the system itself. Some of these systems require highly sensitive data to operate, such as banking, government, and so on. This capturing/inputting of sensitive data can lead to some unintended and serious consequences (for example, hackers, foreign nations, and so on). Ensuring the entire ecosystem is secure is crucial for the success and evolution of the system.

WELTER: The initial cost and investment in control and monitoring systems to support advanced predictive tools.

BARKLEY: Security, trust, regulatory bottlenecks, and continuous improvement.

DANIELS: Data standards and normalizing data for use in our AI models is a challenge. Many of the systems we have designed and deployed are fit-for-purpose systems in one of the 16 sectors (such as financial or chemical systems).

The data within these systems are typically specific and contextualized for the domain practice and often captured in relational databases that are not easily accessible, extensible, or widely available. We have invented many techniques to harvest the data, including application programming interfaces, data hubs, publication-subscription services, extract–transform–load, and many others.

One example is the seemingly simple concept of a part number. A part will typically have many logical attributes or metadata across a heterogeneous system of systems, including supply chain platforms, partner systems, design tools, manufacturing systems, and so on. Maintaining the record of authority, single source of truth, data lineage, governance, and configuration management of the part and associated AI models across distributed heterogenous systems is a challenge.

DEBERRY: The high costs for AI technology, storage and maintenance of the technology, availability of technically trained resources with the necessary expertise, and incompatibility of the existing infrastructure to support the technology.

KEMPER:

- *Power and controls*: The majority of current systems are older and not fully enabled for direct control by computer-run systems or have essentially "cutouts" where people are at a node and directly control the systems, be it from a control room in a facility or the cockpit of some form of vehicle. In addition to dealing with the power (and cooling) associated with AI hardware, the controls to allow the AI to operate and manipulate must be in place.
- *Public support*: Whether we examine the old-school science fiction of Keith Laumer's *Bolo* or Fred Saberhagen's *Berserker* series, a number of the *Star Trek* versions of AI, "Skynet" from the *Terminator* movie franchise, or HBO's updated *Westworld*, the public has decades of culturally solid, if technically flawed, understanding of AI. This results in the public having built-in reservations. It's not "new" to them, and any opinion in place will not be easily displaced by saying, "This is real life and I'm an expert."

  Any time there is discussion of AI control or assistance, you are likely to hear, "Didn't we have a lot of movies telling us why this is a bad idea?" Every time there is a mishap involving AI, whether it's "Trey" (Microsoft's chatbot) or an autonomous vehicle mishap, it plays into existing confirmation bias regarding AI. This situation is more of an issue in Western-style democracies, where public opinion has weight, but this is also where much of the innovation and capital resides. This situation must be factored into management and implementation decisions in a responsible manner, without deception of the public.
- *Systems integration*: There is a disconnect between "traditional engineering" and software development with respect to project management, design responsibility, due diligence, and transparency. Where traditional engineering is oriented toward "it must work," software often is able to "catch it in the next update." In a pure software setting, this may be the case, but, whenever there is integration with hardware of any sort, particularly complex systems, the software must be treated the same as any other system and held the same standards of performance. It doesn't matter if a failure is due to improper brake specifications, a defective motor, or a software error; the resultant failure takes out the whole system.

  A number of recent high-visibility failures, such as the 737 MAX and autonomous vehicle-related fatalities, have highlighted the consequences of the failure to field a fully integrated hardware/software system. AI systems are orders of magnitude more complex, underlining the need for a systematic systems integration consistent with all of the other systems and not dependent on a "we can send a patch" mentality.

***COMPUTER*:** In general, what are the limitations of AI for critical systems?

SCUTA: I have heard there can be biases in some AI engines. Ensuring a non-biased system is key, particularly in a critical system that involves people directly—banking rates, and so on.

WELTER: The availability of applicable failure data to train the system. That's why a lot of the training data come from simulations.

BARKLEY: Systems need to be trained for every possible scenario (or failure condition), visual interpretation like humans (that is, 3D image processing), thinking and learning like humans, and the failure of input sensors.

DANIELS: Value judgments will continue to be part of the coding of any AI model. Data scientists have a responsibility to assert control over the cost, bias, and error functions to ensure the most helpful implementation of any AI algorithm. We must match the need as represented by the domain expert (the user community) against the math/coding required to produce results.

Algorithms routinely experience punishment during their learning process. For example, let's consider reinforcement learning in playing a game, such as Go. When the computer agent puts down a piece and is immediately surrounded and loses the space, it is punished. After a number of scenarios, the computer agent will avoid moving where it is surrounded. It's a simple reward-and-punishment model.

I don't think we are close to achieving this from a general AI standpoint, but, again, within a limited problem space—where most AI applications exist today—I think it may be possible. I could imagine an ML application that is trained on a data set that includes moral aspects within a defined problem space, for example.

DEBERRY: Since every environment is unique, there is not an off-the-shelf solution to fit every organization since AI can only perform in the manner it was programmed to operate. AI does not possess the ability to think nor the creativity to determine what is best for each environment and, thus, requires intervention from multiple resources to be fully functional.

KEMPER: Currently, the limitations are more in terms of a narrow scope of focus and operation. To use a military example, we are comfortable using AI to be the "smart guy" for one thing, such as "track incoming artillery and shoot them out of the sky." We are not comfortable with giving the AI more operational control, such as "plan and control the integrated artillery plan." This is because that puts the AI in a position to make a decision currently requiring multiple people to approve due to the consequence of artillery rounds striking the wrong place. Some of this is computational limitations, some is a desire for overall human control, and some is an inability to connect enough systems to gain the inputs (sensors) and outputs (send commands that respond at computer speed, not human speed) to attempt to do more. All of this is likely to decrease over time.

POPA: One limitation aspect that I always keep in mind is the ability of the models to extrapolate. For example, while artificial neural networks are extremely powerful tools for both classification and prediction, they are notorious interpolators. What this means is that they are only capable of operating within the space and the data where they were trained. Therefore, the occurrence of an outlier that was never thought of, seen, or learned by the model can lead to a flawed outcome, with safety or environmental consequences. This can also be the case with biased data where AI models are only presented with certain classes, leaving other viable outcomes outside of training or learning sets.

***COMPUTER*:** What are some of the biggest security concerns in these kinds of systems?

WELTER: For nuclear power applications, they must be designed so there is no safety concern.

BARKLEY: Physical security; tampering with sensors, data, or the model connected to the world; and communication security.

DANIELS: Trust is essential in our systems. If users do not trust the AI agents, models, and outcomes, they will not use them. Trustworthy computing in securing critical infrastructures, such as data lakes, pipelines, and structures, is one of the biggest cybersecurity concerns.

We are on the cusp of pervasive connectivity with devices, smartphones, vehicles, power systems, medical sensors, tooling, and so on. As we stream telemetry data to build our AI models with greater accuracy, consider the various attack surfaces, such as hybrid deployed systems, distributed and centralized systems, edge computing stacks, production control systems, endpoint services, mobile devices, and many other architectures.

DEBERRY: One concern would be the protection of highly confidential and sensitive data that may need to be shared for power grids to be at optimal performance. Another would be an exposure or breach of a system's algorithms, which would allow "bad actors" to quickly create an automated method to attack.

KEMPER:

- *Spoofing*: Purposely feeding in bad data or intentional patterns to "trick" the AI into a desired learned response. While this was done by humans with the Microsoft "Trey" chatbot, this would be more likely to be an AI/counter-AI interplay between conflicting systems, whether it's nation-states or corporate rivalries. If properly executed, the "learned behavior" would not be triggered under regular conditions that would allow operations and maintenance to detect it and mitigate.
- *Macro control*: This is roughly the same issue as "who has the codes" for nuclear weapons. This would be for an AI developed for a specific purpose, like "take over a foreign energy grin," with some group in charge of deciding whether to activate it or not. The issue of power and control can be potentially analogous to nuclear weapons release authorities, which is a multilayered and compartmentalized system that has, fortunately, never been fully tested to execution.
- *System brittleness*: Once the given system is reliant upon AI to function safely, the AI potentially becomes a potential single point of failure. This is more likely where there is valuable data, such as financial or intelligence systems, than in a "flow control" application, which would allow for a more distributed node system, such as a power grid designed to operate with AI node loss.

POPA: There are so many angles to address this question. I will approach this from the security concerns of systems penetrability by an ill-intended persona. With the advances in hardware speed, hacking programs are becoming smarter and faster in breaking down security protocols. Breaches in critical infrastructure systems would create significant consequences for safety, public health, and the environment. The advances in quantum computing are a game changer. In the wrong hands, I see the capabilities of this new technology as the biggest concern for all critical systems, especially the energy, banking, and medical sectors.

SCUTA: Henry Kissinger used to say, “Control oil and you control nations; control food and you control people.” I believe his quote needs updating: “Control critical systems and you control nations; control data and you control people.” In my opinion, the security of the inputs and outputs of the AI systems are the most vulnerable portion of the system itself—and where most of the efforts by external parties (hackers or nations) are concentrated to bring systems down. So, ensuring there is a major cybersecurity effort evolving in parallel is necessary.

***COMPUTER*:** How do we assure the public of the safety, security, reliability, and so on of critical systems with embedded AI?

BARKLEY: Walk the talk—for example, Elon Musk should be comfortable driving his car in autonomous mode at 80 mi/h. Build the confidence gradually—start with basic modes with human backup and gradually move toward full autonomy. Implement continuous improvements—over the air.

DANIELS: As engineers, we have a responsibility to ensure the safety, security, and reliability of mission-critical systems. It starts with education and continuous learning throughout our practice. In fact, when a professional pursues the professional engineering credential, he or she starts with the professional conduct and ethics exam.

The U.S. Department of Defense (DOD) Joint AI Center (JAIC) released the AI Education Strategy in 2020. The strategy identifies investments in AI education and training that increase the U.S. national AI workforce capacity, bolstering U.S. security and economic competitiveness. We recently updated the Institute of Industrial and Systems Engineers (IISE) *Maynard Handbook* to include AI, data sciences, and the importance of the ethical practice of AI. I am excited to see the inclusion of modern practice in our traditional industrial and systems engineering discipline.

DEBERRY: We must normalize our data, ensure the integrity of the data, and perform rigorous testing to validate and verify that the technology is performing to specifications. In addition, we must maintain ethical, value-aligned, enforceable guidelines for the use of the technology.

KEMPER: Issues of the public’s perception are often rooted in popular imagery. Fundamentally, the public wants to believe that, if a hostile AI agent was trying to take control of systems, such as in the first *Transformer* movie, a general can grab an axe and chop the line needed to defeat the intrusion.

As a technical and security professional, I recognize how problematic the scene was, but it doesn’t change the power of such imagery of a decisive option. How to give the public that level of assurance is problematic because people are not in a position to understand the systems, and, if they were, the applicable security would prevent sharing vulnerabilities, planned or otherwise. For the next generation or two, it is likely there will be a need to have “human override options” in all systems, such that even a fully automated aircraft will have qualified pilots in the cockpit “just in case,” even if they go years without needing to act.

Special training and certification will be helpful. It will be more critical in dealing with allied design and engineering since the training and certification will also include some focus on cross-discipline processes and overall management. This could be seen as “the public” from an AI systems-centric perspective.

The key aspect of including the other engineering and computer disciplines in peer-to-peer relationships of mutual understanding is to not only improve the overall AI implementation but to have these other professionals see the “AI side” as a partner. This is opposed to movie portrayals where the AI is a “black box” dropped in to control systems designed by others, with only a select few understanding and controlling the “scary” AI. If the other disciplines understand its capabilities and limitations, at least to some degree, then this will be likely to filter to the public as a whole.

POPA: We are coming close to the point where embedded AI systems are becoming part of our daily life. In certain cases, the user does not even know that he or she is operating an AI model. While some segments of the population may not be concerned, it is the fiduciary responsibility of the creators to ensure the viability as well as explain the limitations and outcome of the AI models.

We need to start slowly shifting from the black-box concept to explainable models, which are transparent and able to justify/rationalize the outcome. Therefore, prior to deployment, a peer review and audit should be considered to ensure the functionally. With regard to the public, one of the best assurances is continuous education.

SCUTA: Generate awareness—so people realize AI is here for their best interest and not to “dominate humans.” Maybe a certification and some training courses are a good way to start this awareness campaign, particularly in companies that have some interaction (direct or indirect) with a critical system. Such training can ensure that the weakest link of the system becomes stronger, and it can be as simple as making sure members increase of the security/complexity of a password, and so on.

WELTER: We must show that, if these systems fail, they are still safe. Employ defense in depth.

***COMPUTER*:** Are there real ethical dilemmas in AI in critical systems?

DANIELS: We have an ethical responsibility not to allow a critical system to grow beyond our understanding of complexity. We must design in guard rails and cross checks from the earliest stages of development, even at the requirements gathering stage. Learning functions (that is, data exposed to the model in training, covariance checks, and testing) and logic can and should be documented and coded in ways where peer and external/periodic reviews maintain an essential place in the lifecycle of the AI implementation. Human-centered design is a concept espoused at the DOD's JAIC and further manifested in Lockheed Martin's design.

Maintain human oversight and system boundaries. The system should be limited in its problem space and ability to take self-directed action—ensuring human oversight at these boundaries. The AI may make a "wrong" decision (from the human's viewpoint), but the boundaries should seek to limit the negative consequences of such a wrong decision. Even when the decision is based on an extremely complex set of inputs and relationships, we should seek ways for humans to understand the decisions—not in real time, but with some ability to understand why the decision was made and a way to correct for it in the future.

DEBERRY: Yes, as with many technologies, there are limited regulatory controls on the use of AI (for example, North American Electric Reliability Corporation) and the reliance on AI to the point that any failure in the technology could prove catastrophic without proper manual procedures as a contingency plan.

KEMPER: Yes. Going beyond "yes" invites "with respect to what?" To keep it simple, a huge ethics issue will be based on competition. If it is perceived that increasing AI control increases the advantages it provides (for example, faster decision making, a greater depth of data processing, more integrated control of systems, and so on), there will be competition among economic rivals to not only increase power but reduce the messy human aspect.

For example, financial futures trading, which can trigger a crash (or worse) similar to the 2010 "flash crash," can be more given to using AI to eke out that fractional advantage by trading the right thing at the right time and find ways to gain advantage in one area by undermining value elsewhere. If a system owner exerts more control and has more limits on procedures (programmed ethics), the systems with less human intervention and limits could earn more. Earning more, or any other performance outcome the AI is supposed to provide, will put pressure on matching or surpassing others.

Even if you do have more controls or limits than others, it does not mean you have enough. It's a false measure to "not be the worst offender." Take these same concerns and apply them to a worldwide system where some economies are market driven and others are command driven and autocratic, such that a crash will be exploited because popular opinion and accountability are not a driving concern.

Now, take that same premise with different nation-states and apply it to the military. All of the world's major military powers are exploring how to take advantage of crewless autonomous systems as well as AI-powered control systems. If an AI targeting system can "kill" that many more legitimate targets but with more "collateral damage," will an opposing nation stick to a more humane or rule-of-law-compliant system, even if this means you are at a significant disadvantage? If you lose by being more humane, what is the consequence?

It is conceivable that the AI's capabilities to control systems efficiently can get to the point of the same net effects of mutual assured destruction for nuclear devices where, if the systems "go hot," the competing AIs would wipe out all sides with nonnuclear weapons. It would not be like in the movie *War Games*, where a sole AI decides it would be better to play checkers.

The five facets of ethics in AI can be listed as follows:

- *Responsibility*: Who is responsible for AI?
- *Governance*: How is AI controlled?
- *Trust*: How can AI be trusted?
- *Law*: How can AI be used lawfully?
- *Traceability*: How are the actions of AI recorded?

To give a measure of how mainstream the concern is, the following is a continuing professional development course offered to Australian engineers on the topic of ethics and AI: https://engineersaustralia.org.au/event/2021/05/ethics-ai-defence-36781. In addition, it's an "up-front" issue for Nordic Engineers.[5] These issues are mainstream engineering.

It's also not theoretical. It's already in the mainstream media that a drone "in autonomous mode" has targeted and killed a person based on electromagnetic emissions.[6] All one has to do is extrapolate that a critical defense system would use "swarms" of such drones hardened against disruptive emissions and seeking targets that fit a profile. How carefully crafted that profile is becomes an issue. It also raises the question of what matters—the means or the outcome?

In this example, it could be whether the profile was generated by the AI based on the most-current intelligence (theoretically) right before launching from a "mother ship" or was something a human programmed weeks or months prior based on historical data. It would seem that an AI could craft a very careful targeting profile and upload it in flight right before launching the swarm. However, the tolerances on that profile may be loose enough to ensure at least one drone gets a high-value target at the risk of several "almost exactly right," or the profile could be tightened so that only the right person is targeted, but the target can also be missed due to minor variances.

Would an AI-generated "bad target" be different than if a human did the same programming or pointed at a screen and said "shoot"? In theory, yes, because you could fire or (if military) court-martial a human. Can you punish "a toaster"? Who is to blame if the "toaster" kills the wrong person?

POPA: Again, there are many angles from which one can look at this. I would relate this to the intent of the AI model and creator. As an example, one can think of the data collection for all of the personal devices that members of the public have welcomed into their daily lives—smart watches, smart home devices, security systems, and so on. The use of the data that are gathered can be an ethical dilemma when they are used without consent.

SCUTA: Yes, in the end, someone coded part of the system, and one must be careful not to generate a biased AI engine. I believe this happens involuntarily; you don't know what you don't know (Johari's window). This is a hot topic, especially when considering using AI systems for the security of a nation—a U.S. immigration-based AI code could be very different than a Brazilian immigration AI code due to each nation's bias.

WELTER: None yet, really. AI is not real artificial intelligence—mostly ML applications.

BARKLEY: In health care and military systems, ethics do matter.

***COMPUTER*:** What should we watch for with respect to AI in critical systems going forward?

DEBERRY: The ability to better balance the demand and supply of power based upon predictive data, the detection and prevention of utility theft, and the ability to predict power failures. Consumers will be able to determine off-peak hours to tailor energy usage based on times offering the lowest cost.

KEMPER: Systems integration would be an immediate issue, as this is already in play with regular software/hardware challenges. This is not about a Skynet-type takeover but, rather, a 737 MAX type, with the system failing in a predictable manner given uncommon but normal operations parameters. Some of this may be mitigated with human intervention on standby, but that is a planned mitigation for an understood contingency.

What is likely to be more pernicious is something happening that could only reasonably occur if it is AI enabled—something that has not occurred before and is not in the list of things that people agree to plan for. A failure that has a longer system reaction time due to novelty will have more opportunity to cause harm. Systems integration testing, including using stochastic parameters instead of single deterministic sets, will help mitigate the potential frequency and severity of such incidents.

Another thing to watch for is the opposite of using deliberate systems integration—an "arms race" or "space race" competition that rewards cutting corners that do not result in immediate failure. This could be within a nongovernmental sector like finance or in developing crewless autonomous systems for the military.

Ethics matter the most when they are the least convenient. The more there is a deadline or "we are going to lose" pressure or it becomes a choice of "will we do this to survive," the more ethics matter. Be aware that there is a solid logic behind "if we have to go into an elevated risk mode if it's about survival (as a company or a nation), then it means we should be willing to go into the same mode in order to keep us from being in a life-and-death decision in the first place." In other words, it's the same logic of a preemptive strike to stop a threat before it is active.

Another key point is that the person with ultimate executive power over such decisions is often not a technical specialist in a given area. This is not to say "we are cursed to be ruled by the ignorant"—it is just reality that a person in charge of a government or large organization will only have so much technical expertise with respect to the group as a whole. You cannot know everything. The issue is that the role of ethics in AI must be treated as an engineering-wide—if not a society-wide—issue to have those principles influence decision makers at all levels rather than being limited to the AI specialists with the unenviable duty of being the sole gatekeepers.

POPA: To quote the Dario Gil from IBM, we are at an inflection point where our "narrow" form of AI systems have begun to work and been accepted within our society for very specific tasks and objectives. This would include image and speech recognition, common autonomous tasks, and so on.

The next phase is what is called the *broader AI*. What that means is leveraging what we learned from the existing AI model within "narrow" AI and expanding without starting again from scratch. That would lead to a broad AI era. This kind of thinking for AI models and applications will soon be reflected in the critical systems.

SCUTA: Security, security, and security. I believe that the integration of AI into critical systems is at an extremely high pace. However, I am not so sure the security around these systems is evolving at the same rate. The private sector tends to be better at keeping this disparity at bay, but the public sectors typically only change/update after a disaster happens (push rather than pull). Generating awareness, educating the public and private sectors, and allocating needed resources are key parts of maintaining our critical systems as operational and free of external threats.

BARKLEY: Systems that are backed up by humans just in case of a system failure or unrecoverable situation. Proactive learning (training) and not being dependent on reactive learning after the disaster. Continuous improvements.

DANIELS: We are at an inflection point where AI-enabled critical systems are becoming feasible and accepted. Explainable AI will be key to understanding what the data are telling us and how responsively our critical systems will behave. I am optimistic about the future of AI, and I continue to partner with the JAIC, IISE, and other professionals to extend the ethical use of AI and maintain a leadership position for our country.

Kevin Scott, CTO of Microsoft, recognized that

> *. . . [when] we invented software engineering as a discipline over the course*

of the past 60 years or so, we realized that finding all of the bugs in software is hard. We built a whole bunch of practices to try to catch the most common type of software bugs. We created a set of techniques to help us mitigate the impact that the bugs that slip through will have. We're going to have to build a similar set of things for machine learning models and AI.[2]

Our panel of experts agreed with this observation and provided many recommendations on how these assurances might be obtained. ■

REFERENCES

1. "Critical infrastructure sectors," Cybersecurity & Infrastrcuture Security Agency, CISA, Arlington, VA. https://www.cisa.gov/critical-infrastructure-sectors (accessed 23 June 2021)
2. K. Scott, "Forward thinking on artificial intelligence with Microsoft CTO Kevin Scott," McKinsey Global Inst., New York, podcast, June 10, 2021. https://www.mckinsey.com/featured-insights/future-of-work/forward-thinking-on-artificial-intelligence-with-microsoft-cto-kevin-scott (accessed June 23, 2021)
3. M. Gladwell, *Blink: The Power of Thinking Without Thinking*. Little Brown, New York: Back Bay Books, 2005.
4. K. Griffith, "Autonomous corridor plan moves forward in Maryland," Government Technology, May 21, 2021. [Online]. Available: https://www.govtech.com/fs/autonomous-corridor-plan-moves-forward-in-maryland
5. "AI & ethics," Association of Nordic Engineers, Copenhagen, Denmark. Accessed: May 31, 2021. [Online]. Available: https://nordicengineers.org/artificial-intelligence/
6. Z. Kallenborn, "Was a flying killer robot used in Libya? Quite possibly," Bulletin of the Atomic Scientists, May 20, 2021. [Online]. Available: https://thebulletin.org/2021/05/was-a-flying-killer-robot-used-in-libya-quite-possibly/

PHIL LAPLANTE is professor of software and systems engineering at Penn State, Malvern, Pennsylvania, 19355, USA, and an associate editor of *Computer*. Contact him at plaplante@psu.edu.

BEN AMABA is the CTO for digital transformation with technology innovation and execution at IBM, Collierville,Tennessee, 38017, USA. Contact him at bamaba@us.ibm.com.

Rejuvenating Binary Executables • Visual Privacy Protection • Communications Jamming
IEEE SECURITY & PRIVACY
Building Dependability, Reliability, and Trust
SOFTWARE EVERYWHERE
January/February 2016 Vol. 14, No. 1
Policing Privacy • Dynamic Cloud Certification • Security for High-Risk Users
IEEE Symposium on Security and Privacy
March/April 2016 Vol. 14, No. 2
Smart TVs • Code Obfuscation • The Future of Trust
ECONOMICS OF CYBERSECURITY PART 2
May/June 2016 Vol. 14, No. 3
IEEE
IEEE computer society CELEBRATING 70 YEARS
Reliability Society

Digital Object Identifier 10.1109/MC.2021.3109699