Reducing Human and AI Risk in Autonomous Systems

On two successive days, 17 and 18 September 2021, *The New York Times* published articles that seemed to question which was superior with respect to weapons systems: human decision-making or artificial intelligence (AI). The first article describes how a series of human errors led to a drone, controlled by the US military, firing a missile at a friendly target. The target was mistakenly thought to be a vehicle driven by ISIS-K terrorists containing explosives and headed for Hamid Karzai International Airport in Kabul, Afghanistan.[1] The second article describes how Israeli agents used face-recognition technology to identify and kill an Iranian nuclear scientist with amazing precision using a remote-controlled machine gun, while sparing his wife, who was sitting in the vehicle next to him.[2] What do these two examples say about human-controlled vs. autonomous cyberphysical systems operating weapons, self-driving cars and the like? Although they are indicative, they are far from representative of the whole issue. Nevertheless, they bring to the fore a number of questions regarding the thought processes, biases, ethics and trustworthiness of human, computer-assisted, automatic and autonomous systems, particularly weapons systems. In addition, there are concerns about safety when life-or-death decisions and actions are transferred from humans to machines. Some fear that AI-based autonomous systems might exhibit artificial general intelligence (AGI) surpassing human intelligence, representing an existential threat to humankind.

To mitigate risk attributed to semiautonomous and fully autonomous systems, it is important to understand the differences between how human brains work and how computer AI systems function.

C. WARREN AXELROD | PH.D., CISM, CISSP

Is the research director for financial services with the US Cyber Consequences Unit. Previously, he was the business information security officer and chief privacy officer for US Trust. He is a cofounder and served as a board member of the Financial Services Information Sharing and Analysis Center (FS-ISAC) and represented the banking sector's cybersecurity interests in Washington, DC, USA, during the Y2K date rollover. Axelrod received the ISACA® Michael P. Cangemi Best Book/Article Award in 2009 for his *ISACA® Journal* article "Accounting for Value and Uncertainty in Security Metrics." He was honored in 2007 with the Information Security Executive Luminary Leadership Award and received a Computerworld Premier 100 award in 2003. In addition to authoring the books *Engineering Safe and Secure Software Systems* and *Outsourcing Information Security,* Axelrod edited *Enterprise Information Security and Privacy* and has published more than 140 professional articles and chapters in books. He has delivered more than 150 professional presentations. His current research includes the behavioral aspects of cybersecurity risk management and the security and safety of cyberphysical systems, particularly as they relate to autonomous road vehicles.

As computer systems become less reliant on human decision-making, they assume more control, although some residual human intelligence (e.g., biases, ethics, trust) is likely to exist in even the most independent AI systems.

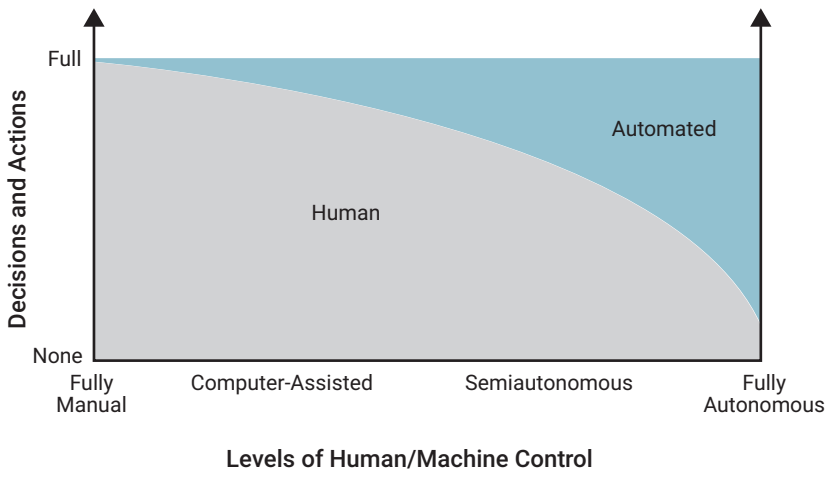
To mitigate risk attributed to semiautonomous and fully autonomous systems, it is important to understand the differences between how human brains work and how computer AI systems function. Then it is necessary to ensure that autonomous systems are designed and built to allay fears by considering biases, ethics, fairness and trustworthiness as they apply to AI systems.

Human vs. Machine Decisions and Actions

Figure 1 shows how the ratio of human to machine decision-making and actions diminishes as control moves from manual through computer-assisted to autonomous (when computer systems assume almost complete control). As computer systems become less reliant on human decision-making, they assume more control, although some residual human intelligence (e.g., biases, ethics, trust) is likely to exist in even the most independent AI systems.

FIGURE 1

Human vs. Machine Decisions and Actions at Various Levels of Autonomy



The idea of complete autonomous takeover is approached with skepticism. For example, an Institute of Electrical and Electronic Engineers (IEEE) article noted, "The U.S. Army is particularly wary of relying on black-box systems, so Army researchers are investigating a variety of hybrid approaches to drive their robots and autonomous vehicles."[3] This same concern is voiced with regard to self-driving cars reaching Society of Automotive Engineers (SAE) level 5, or full automation, which means that the vehicle performs all driving tasks under all conditions with no human attention or interaction.[4]

Although the shape of the curve dividing human and machine involvement is speculative, one might imagine a tipping point at which decision-making rapidly shifts from predominantly human control to AI control. Whether this will result in a monumental change depends on the differences between human and AI thought processes and values and whether such differences are adequately understood and accounted for in advance. If they are, the transition will be relatively painless. If they are not, there may be some unpleasant surprises. Therefore, it is imperative to examine the different ways humans and machines think, analyze, approach and respond to a full range of situations, and to allow for such differences when designing, creating and operating future systems.

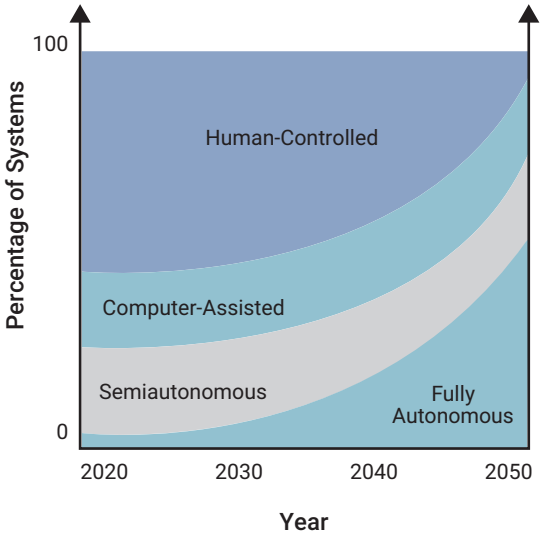
Nevertheless, in even the most autonomous systems, some measure of human intervention can be expected. Problems experienced with fully autonomous self-driving cars demonstrate how difficult it is for drivers to relinquish all control to the AI system.

In the 1968 movie *2001: A Space Odyssey*, the HAL 9000 computer, which runs the spaceship, takes over and murders the entire crew except for one astronaut. The surviving human is able to deactivate the computer by removing circuit boards inside the machine.[5] This suggests a critical need for an off or cancel switch in all autonomous systems.

One issue is how to determine the optimal levels of automation and autonomy for particular applications, recognizing that technologies and situations are constantly changing. Currently, there is justifiable reluctance to hand over complete control to AI systems, although some are pushing to move ahead regardless. **Figure 2** shows how the mix of manual and autonomous systems might change over the next several decades. Although it is not based on any specific forecasts, it indicates how systems are evolving over time.[6]

FIGURE 2

Changes in the Mix of Systems Over Time



These projections are based on the presumption that the growth in AI systems will continue on an upward trajectory, as illustrated in **figure 3**. However, two prior cycles consisted of rapid growth followed by AI winter, a dormant period during which there was a decline in interest in AI. Because AI technology has been designated a US national security requirement, such declines are unlikely in the future.[7] It is reasonable to assume that growth in AI will accelerate, leading to a rapid increase in the relative percentage of autonomous systems.

Automatic, Intelligent and Autonomous AI Systems

Distinguishing between automatic, intelligent and autonomous systems is important, especially as any system can belong to more than one category. Indeed, all autonomous systems are intelligent and automatic, but all automatic systems are not intelligent or autonomous. This differentiation is relevant because as systems are combined into more complex systems-of-systems, their components must be identified to achieve the transparency and understanding needed to manage them.

Here, automatic systems are defined as those that operate in a specific predetermined manner when a particular trigger or button is activated. Intelligent systems are those that, once activated, respond differently to various stimuli. Activated autonomous systems are those that operate without predetermination or intervention. For example, an automatic rifle responds to pressure on its trigger, whereas a vehicle's automatic transmission is actually an intelligent automatic system that takes in data from a number of sensors and uses algorithms to determine when to change gears. Autonomous systems take in external data, but their responses are not predetermined; they adapt to various circumstances and respond to changing situations in the sense that they learn from prior experience—referred to as machine learning (ML) and deep learning—and act accordingly.

Human-implemented automatic systems—namely, those that operate by themselves without the initiator's direct intervention—likely go back eons. The first type of automatic system may have been an animal trap, which is activated when an animal steps on a trigger. Land mines operate in the same way. Modern-day automation got into full swing with the Industrial Revolution as water-driven and steam-driven engines replaced manual labor. The main characteristic of automatic machines is that their actions and responses have been predetermined and, barring faulty design, malfunction or failure, such systems operate time after time according to the intentions of their designers. For computerized automatic systems, the requirements and specifications are established during early design phases, and subsequent verification and validation phases ensure that the product or system meets those requirements.[8]

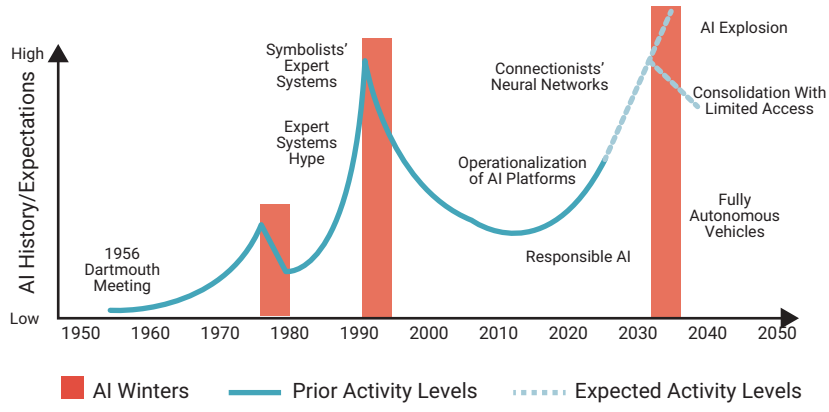
Intelligent systems are generally endowed with analysis and reporting capabilities, which sometimes extend to some form of action. Semiautonomous systems take intelligent systems a step further in that they are guided by humans to a certain

LOOKING FOR MORE?

- Read *Auditor's Guide to Machine Learning Part 1, Technology www.isaca.org/audit-practitioner-guide-to-ML-part-1*
- Learn more about, discuss and collaborate on risk management in ISACA's Online Forums. *https://engage.isaca.org/onlineforums*

FIGURE 3

History and Expectations of AI-Based Systems



Sources: Brooks, R.; "A Human in the Loop: AI Won't Surpass Human Intelligence Anytime Soon," *IEEE Spectrum*, vol. 8, iss. 10, October 2021, *https://ieeexplore.ieee.org/document/9563963*; Strickland, E.; "The Turbulent Past and Uncertain Future of AI: Is There a Way Out of AI's Boom and Bust Cycle?" *IEEE Spectrum*, vol. 58, iss. 10, October 2021, *https://spectrum.ieee.org/history-of-ai*; Panel for the Future of Science and Technology, "The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence," June 2020, *https://www.europarl.europa.eu/stoa/en/≠document/EPRS_STU(2020)641530*

FIGURE 4

System Categories, Subcategories and Areas of Concern

Category of System	Subcategory of System	Examples of System	Concerns					
			Security	Safety	Bias	Fairness	Transparency/ Explainability	Trustability
Automatic	Decision-Making	Expert systems	X				X	X
	Physical	Vehicle auto transmission		X				X
	Cyberphysical	Industrial robot		X			X	X
Intelligent	Decision-Making	Chess playing	X			X		X
	Physical	Lane correction	X	X	X			X
	Cyberphysical	Internet of Things (IoT) industrial controls	X	X	X	X		X
Autonomous	Decision-Making	Programmed trading	X		X	X	X	X
	Physical	Standalone robots		X	X		X	X
	Cyberphysical	Self-driving cars and weapons	X	X	X	X	X	X

FIGURE 5

Primary Functions of Regions of the Human Brain

Brain Region	Primary Functions
Frontal lobe	• Problem-solving • Emotional traits • Reasoning (judgment) • Speaking • Voluntary motor activity
Parietal lobe	• Knowing right from left • Sensation • Reading • Body orientation
Temporal lobe	• Understanding language • Behavior • Memory • Hearing
Occipital lobe	• Vision • Color perception
Cerebellum	• Balance • Coordination and control of voluntary movement • Fine muscle control
Brain stem	Nonvoluntary activities: • Breathing • Body temperature • Digestion • Alertness/sleep • Swallowing

degree, beyond which they act on their own. Fully autonomous systems encompass the full range of detection, analysis and response.

Figure 4 gives examples of the types of systems falling into the automatic, intelligent and autonomous categories and their subcategories, and highlights some of the concerns applicable to those systems.

Human vs. Machine Thought Processes

There are many functions of the human brain (**figure 5**), and some may be replicated in computer systems (**figure 6**).

Although a great deal is known about the workings of the human brain, some areas remain mysterious, such as the cerebellum, which is situated below the main cerebral cortex at the back of the skull. Given its size, the cerebellum has a much greater surface area compared with the rest of the brain, due to the presence of more ridges than in the cerebrum, and it is generally thought to coordinate and control movement. However, research indicates that it may have a cognitive function as well.[9]

Since designers and developers of AI systems tend to ignore the functionality of the cerebellum, it could be the site of functions that bridge the gap between brains and their electronic emulators. If the cerebellum is proved to have specific cognitive

FIGURE 6

Replication of Brain Functions in Computer Systems

Brain Area	Machine Equivalents	Comments
Frontal lobe	Computer systems ranging from those with predetermined activities through expert systems to machine learning and AI, including cyberphysical systems and robotic systems	Emotional traits may be emulated to some extent using AI, but the traits are not intrinsic to such systems.
Parietal lobe	Industrial and personal control systems such as robots, autonomous and semiautonomous vehicles and IoT devices	Sensors are components within cyberphysical systems, such as industrial control systems and robots, that provide inputs and feedback to control systems.
Temporal lobe	Translation and speech recognition systems, for example, which are rule-based but generally do not have any understanding of linguistic premises	Computer systems can emulate some temporal lobe functions and appear to have the ability to understand, but they do not actually have that ability; understanding results from answers to "why" questions.[a]
Occipital lobe	Face-recognition and robotic systems	There are many privacy and ethical questions as to the use of facial recognition and concerns about biases introduced in caching learning form selected sample data.
Cerebellum	Systems that monitor, review and coordinate motor functions and use feedback to correct discrepancies, such as systems that control mobile robots	New research indicates a possible role of the cerebellum in cognition.
Brain stem	Automatic control and operating systems for computers, networks, devices and heating and cooling systems	Functions relate to autonomous and semiautonomous activities.

Source: a) Ackoff, R. L.; *Ackoff's Best: His Classic Writing on Management*, John Wiley and Sons, Inc., USA, 1999, *https://www.wiley.com/en-us/Ackoff%27s+Best%3A+His+Classic+Writings+on+Management-p-9780471316343*

functions, it might explain why current models of the brain lack significant components. This could also account for some of the differences between human intelligence and AI.

With regard to the better-understood brain functions, it is clear that thinking-related functionality is distributed among the various lobes, whereas automatic functions, such as breathing and digestion, are concentrated in the brain stem. In a sense, the same is true of computer systems. The processing aspects are written into application programs and the basic housekeeping functions reside in system programs and firmware.

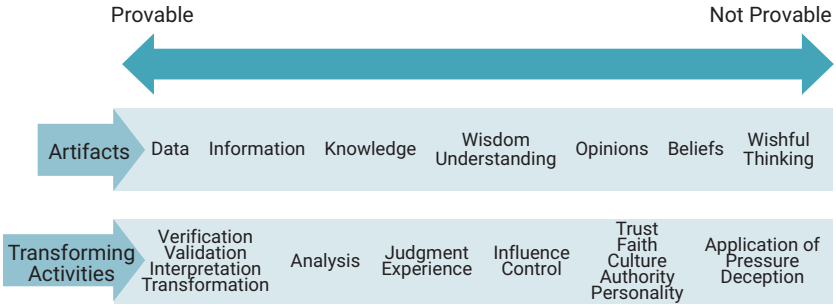
Many computer systems emulate limited processes within a single lobe of the brain, although some combine the functions of different lobes, such as a device that listens in one language (temporal lobe), translates into another language (temporal lobe) and then speaks the translated phrase (frontal lobe). AI systems, in contrast, often combine multiple processes across various lobes in a complex manner, although they still tend to concentrate on specific subsets of brain functionality. Comparisons between organic brain processes and body functions and those composed of software, electronics and mechanical systems can be problematic when the latter are intended to replace the former, particularly in the area of cognition.

Machines lack human understanding (e.g., answers to "why" questions, conveyed by explanations), emotions (e.g., fear, anger, happiness, regret) and motivations (e.g., greed, survival, security, safety, ego satisfaction), even when they have humanoid features and respond in ways that might be expected of humans. Only the designers and creators of these machines have emotions and motivations; such sensitivities are not native to the machines or to the software programs themselves. However, machines' actions do reflect specific motives or intent (rather than motivations) to the extent that they do what is expected of them. Conversely, automatic machines generally follow specific sets of preprogrammed rules, except when design or code errors lead to machine malfunction or failure in the field. One must be particularly sensitive to system behavior in a laboratory setting vs. in the field. A system that operates as intended in a controlled environment, such as a development facility, may behave otherwise—to the extent of failure—when deployed in an operational situation. It is important to conduct final tests in real-world environments.

Figure 7 shows a spectrum of human thinking, ranging from provable facts to beliefs and hopes that cannot be proved. Human thinking is influenced by experience and emotion, and as a result, people interpret information in specific ways, depending on their backgrounds and cultures. Data become

FIGURE 7

Provability Spectrum of Human Thinking



Source: Adapted from Ackoff, R. L.; *Ackoff's Best: His Classic Writing on Management*, John Wiley and Sons, Inc., USA, 1999, *https://www.wiley.com/en-us/Ackoff%27s+Best%3A+His+Classic+Writings+on+Management-p-9780471316343*

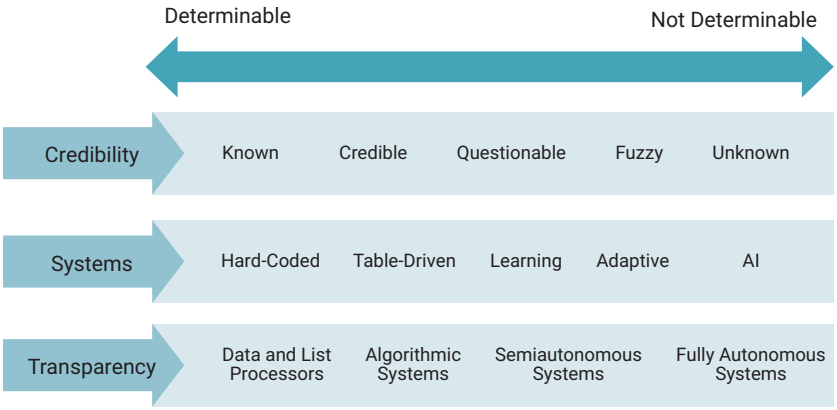
information as they are interpreted. Information is transformed into knowledge through analysis, and knowledge becomes wisdom when judgment is applied. However, opinions and beliefs take over from wisdom when the balance shifts due to influences such as culture and authority. Beyond that is the realm of wishful thinking, where it becomes a matter of wanting to believe (but perhaps not actually believing) in fundamentally deceptive information, which might be neutral, as in misinformation, or damaging, as in disinformation. As one moves from data to wishful thinking, reliance on facts diminishes, and opinions and beliefs begin to dominate. This does not mean that beliefs are superior to wisdom. Here, the progression from data-based activities to wishful thinking is expressed in terms of provability, and wisdom is usually more provable than beliefs.

In contrast, computer systems—whether software, machines or a combination of both—exhibit a somewhat different spectrum of thinking (**figure 8**). This is because, if computer systems are programmed correctly, they do what they are told to do and do not subject information to human considerations, unlike AI systems, which may be taught to interpret information in a quasi-human manner. Rather than provability, computer systems operate under the expectation of determinability. That is, they generally do not weigh in on whether information is provable unless that is their specific purpose. There is a difference between AI and autonomous designations. Although AI underlies and enables autonomous systems in general, many AI systems, such as decision-making systems, are not autonomous and require human involvement.

These different ways of thinking between humans and computer systems explain why it is so difficult to replicate human thinking in intelligent systems.

FIGURE 8

Spectrum of Computer System Thinking



Although AI underlies and enables autonomous systems in general, many AI systems, such as decision-making systems, are not autonomous and require human involvement.

Bias, Fairness, Trustability, Transparency, Explainability and Ethics

The main differences in the approaches of humans and computers can be examined through their relative biases, fairness, trustability, explainability and ethics.[10] Human biases and prejudices are preconceived beliefs or opinions that are not based on reason or experience. System biases are introduced by humans at various points in a system's life cycle; systems do not have their own beliefs or opinions. System biases may arise intentionally or inadvertently, either with or without the knowledge of the humans involved. For example, data used in the design and testing of systems may be intrinsically biased in favor of certain ethnic groups, but this may not be realized until the system is operational, if then. Similarly, system ethics reflect the ethics of designers and operators, which vary based on culture, race and background. Ethics may be introduced intentionally or unintentionally. Trustability depends on factors such as knowing all system sources (i.e., provenance) and positive results from authorized testing organizations. Sometimes such trust is misplaced, especially when the system's provenance is not completely known.

Figure 9 considers various aspects of these factors for different system types.

As shown in **figure 9**, less autonomous systems are more passive than more autonomous systems; they tend to exhibit less bias than fully autonomous systems because they incorporate the human element to a greater degree, and humans have different perspectives and come to different conclusions based on how much control they have.[11]

The more autonomous a system is, the less influence humans have on its operation. For example, sensing, monitoring and reporting systems that are not making any decisions do not incorporate ethical issues, whereas ethics can have major significance in a fully autonomous system, especially if it is confronted with life-or-death decisions.

Trustability

When it comes to trustability, less autonomous systems may be more trustable than fully autonomous systems.

This is because of their relative transparency. One can normally inspect the requirements, specifications and code to test the behavior of monitoring and decision-making systems and determine what they are supposed to do and whether they actually do it. As systems become more autonomous and complex, it becomes more difficult to predict how they might behave under different conditions, so that such systems cannot be trusted at the same level as less autonomous systems. There have been attempts to determine how trustworthy AI and autonomous systems are, but the results are not encouraging.[12] This is becoming more of an issue as fully autonomous systems become more complex and their workings more obscure.

As systems become more autonomous and complex, it becomes more difficult to predict how they might behave under different conditions, so that such systems cannot be trusted at the same level as less autonomous systems.

Trustworthiness has come to mean "a set of (overlapping) properties," as indicated:[13]

- **Reliability**—Does the system do the right thing (validation)?
- **Safety**—Does the system do no harm?
- **Security**—How vulnerable is the system to attack?
- **Privacy**—Does the system protect a person's identity and data?
- **Availability**—Is the system up and ready when it is needed?
- **Usability**—Can a human use it easily?

These properties are quite similar to so-called nonfunctional software characteristics.[14] In addition, this definition of reliability coincides with what is determined in the validation process,[15] whereas reliability is generally held to be related to availability. The validation process is used to check that the completed system meets specified user requirements, which is different from the verification

FIGURE 9

Biases, Ethics and Trustability for Various Systems

Type of System	System Passivity	System Biases	Introduced Ethics	Trustability
Monitoring and reporting	Passive	From machine learning and sensor design and ability	Limited	Considerable
Decision-making	• Active for decisions • Passive for actions	From system design creation, testing and data quality	Sometimes	Broad
Human-machine hybrid	Depends on extent of roles of systems and humans	From human biases and system and data selection	Usually required	Somewhat
Semiautonomous	• Active • Some human intervention	From human biases and system and data selection	Some human influence	Limited
Fully autonomous	• Active • No human intervention	Considerable potential for biases	Little human influence	Minimal

process, which ensures that the programmed system satisfies the design.[16] Indeed, this meaning of reliability is more in line with effectiveness.[17] In any event, the listed properties are long-desired attributes of availability from the user's point of view, as opposed to the provider's perspective.[18, 19]

There are other desirable properties required for being able to trust AI systems, including:[20]

- **Accuracy**—How well does the AI system do with new (unseen) data compared with the data on which it was trained and tested?
- **Robustness**—How sensitive is the system's outcome to a change in input?
- **Fairness**—Are system outcomes unbiased?
- **Accountability**—Who or what is responsible for the system's outcome?
- **Transparency**—Is it clear to an external observer how the system's outcome was produced?
- **Interpretability/explainability**—Can the system's outcome be justified with an explanation that a human can understand or one that is meaningful to the end user?
- **Ethicality**—Were the data collected in an ethical manner? Will the system's outcome be used in an ethical manner?

Bias and Fairness

Humans exhibit a wide range of biases, some of which affect the types of behavior system designers introduce into AI/ML computer systems. Designers and developers who create intelligent systems may or may not be aware that they are introducing biases and what their effects might be. With AI systems, biases can be introduced at many stages, from choosing the data on which to perform ML to creating the algorithms that perform assigned functions of the delivered systems. **Figure 10** describes human biases and how AI systems might reflect those biases.

FIGURE 10

Biases and Their System Equivalents

Human Biases	Description of Human Biases and Tendencies	System Representations and Equivalents
Myopia	Focus on overly short future time horizons	Machine logic should not be anticipatory over short or long terms.
Recency	Cognitive bias that favors recent events over historic ones and gives greater importance to the most recent event	This is formalized by inclusion of such statistical methods as exponential smoothing in algorithms.
Primacy	Cognitive bias that favors items at the beginning (of a list) vs. items that come later	Machine logic should be neutral with respect to sequence of items, if so programmed.
Narrative	Tendency toward the imprinting of dramatic story lines	Search algorithms, such as Google, favor items that receive the most attention and place them higher in their search results.
Amnesia	Tendency to forget too quickly the lessons of past disasters	Adaptive systems learn from experience and retain lessons indefinitely.
Optimism	Underestimation of likelihood that losses occur from future hazards	Machine logic should be neutral with respect to estimating losses.
Inertia	Maintenance of status quo or adoption of a default option	This is a characteristic of machine logic.
Simplification	Selective attendance to only a subset of the relevant factors	Machine logic should usually attend to all relevant factors as programmed.
Herding	Tendency to base choices on the observed action of others	This is a feature of swarm robotics, for example.
Availability	Estimation of likelihood of a specific event occurring based on experience	Characteristic of machine learning.
Compounding	Focus on low probability of an adverse event in the immediate future rather than on the relatively higher probability over a longer time period	This depends on what is programmed into the system or what results derive from machine learning.
Anchoring	Tendency to be overly influenced by short-term considerations that come easily to mind	This is not generally subject to making decision subject to ease of computing or recall.

FIGURE 11

System Biases Introduced During Phases of the AI/ML Pipeline

Phases of the AI/ML Pipeline	System Biases	Descriptions and Causes of System Biases Introduced by Humans
Data creation	Sampling	Due to the selection of particular types of instances more than others, renders the data set under representative of the real world
	Measurement	Introduced by errors in human measurement or because of intrinsic habits of those capturing data
	Label	Associated with inconsistencies in the data-labeling process due to labelers' different styles and preferences or their belonging to different organizational units
	Negative set	Introduced as a consequence of not having enough samples representative of the rest of the world
Problem formulation	Framing effect	Based on how the problem is formulated and how information is presented
Data analysis	Sample selection	Introduced by the selection of individuals, groups or data for analysis in such a way that the samples are not representative of the population intended to be analyzed
	Confounding	Arises if the algorithm learns the wrong relationships by not considering all the information in the data
	Design-related	Solely introduced or added by the algorithm
Validation and testing	Human evaluation	Due to such phenomena as confirmation bias, peak-end effect, prior beliefs (e.g., culture) and how much information can be recalled (recall bias)
	Sample treatment	Introduced in the process of selectively subjecting some sets of people to a type of treatment
	Validation and test dataset	Introduced from sample selection or label biases in the test and validation datasets or from the selection of inappropriate benchmarks and datasets for testing

Source: Adapted from Srinivasan, R.; A. Chander; "Biases in AI Systems," *Communications of the ACM*, vol. 64, iss. 8, August 2021, p. 44–49, *https://cacm.acm.org/magazines/2021/8/254310-biases-in-ai-systems/abstract*

Ideally, AI systems should not exhibit biases that stem from human deficiencies, such as being too selective, being overly optimistic (or pessimistic), exaggerating recent events (unless programmed to do so) and favoring items at the beginning of a list. Biases might also be introduced into AI systems through compounding bias (i.e., focusing on recent adverse events rather than more dangerous long-term developments), anchoring bias (i.e., being influenced by short-term considerations) and availability bias (i.e., estimating likelihoods based on experience).

A major problem is that biases inserted into AI systems, through either the ML process or the programming of algorithms, may not be readily discernible due to a lack of transparency and explainability. This can lead to unfortunate results and poor decision-making, such as accepting or rejecting candidates based on race or gender, due to limitations of the data used in the ML process.

Transparency and Explainability

Because AI systems are so complex, it is particularly difficult for humans (even the system's designers and developers) to understand what is going on under the hood. **Figure 11** describes system biases that can be introduced at each phase of the AI/ML pipeline.[21] Despite attempts to minimize such biases, it is nearly impossible to determine which biases are in effect by observing the outputs and outcomes of the system. Available testing and assurance processes are often inadequate to the task, as one cannot anticipate an adequate range of use cases for systems whose behavior is unpredictable.

Conclusion

As AI evolves, it is important to understand how the human mind works and which characteristics are being transferred—successfully or otherwise—to AI systems. Attempts to emulate human thought processes, emotions and motivations are hampered by the intrinsic differences between how the human brain works and how automated systems

operate. Biases lead to discrepancies between ideal systems and those that are actually produced. Only by examining processes and biases can AI systems be developed that meet operational and ethical requirements. Future research will lead to transparency and a greater understanding of the inner workings of these systems that will eventually dominate people's lives.

Acknowledgment

The author thanks Sam H. DeKay, Ph.D., for his many valuable comments and contributions.

Endnotes

1 Schmitt, E.; H. Cooper; "Pentagon Acknowledges Aug. 29 Drone Strike in Afghanistan Was a Tragic Mistake That Killed 10 Civilians," *The New York Times*, 17 September 2021, *https://www.nytimes.com/2021/09/17/us/politics/pentagon-drone-strike-afghanistan.html?searchresultposition=1*

2 Bergman, R.; F. Fassihi; "The Scientist and the A.I.-Assisted, Remote-Control Killing Machine," *The New York Times*, 18 September 2021, *https://www.nytimes.com/2021/09/18/world/middleeast/iran-nuclear-fakhrizadeh-assassination-israel.html?searchresultposition=1*

3 Strickland, E.; "The Turbulent Past and Uncertain Future of AI," *IEEE Spectrum*, 30 September 2021, *https://spectrum.ieee.org/history-of-ai*

4 Axelrod, C. W.; "The Demise of Self-Driving Cars as Such," Security Boulevard, 20 September 2021, *https://securityboulevard.com/2021/09/the-demise-of-self-driving-cars-as-such/*

5 Kubrick, S., director; *2001: A Space Odyssey*, Metro-Goldwyn-Mayer, USA, 1968

6 There are a number of published forecasts, predominantly in the vehicle and weapons system sectors. See, for example, GlobeNewsire, "The Military Robotic and Autonomous Systems Market Is Anticipated to Grow at a CAGR of 20.75 Percent Based on Market Value During the Forecast Period 2020–2025," 5 November 2020, *https://www.globenewswire.com/news-release/2020/11/05/2120762/0/en/the-military-robotic-and-autonomous-systems-market-is-anticipated-to-grow-at-a-cagr-of-20-75-based-on-market-value-during-the-forecast-period-2020-2025.html.* However, growth is dependent on the acceptance of AI systems, which is expected but by no means guaranteed.

7 National Security Commission on Artificial Intelligence (NSCAI), *Final Report*, USA, 2021, *https://nscai.wpenginepowered.com/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf*

8 Axelrod, C. W.; *Engineering Safe and Secure Software Systems*, Artech House, USA, 2012

9 Wagner, M. J. *et al*.; "Cerebellar Granule Cells Encode the Expectation of Reward," *Nature*, vol. 544, 2017, p. 96–100, *https://www.nature.com/articles/nature21726*

10 Here, the term "trustability" (i.e., able to be trusted) is preferable to the more common term "trustworthiness" (i.e., deserving of trust) because one cannot determine whether a system deserves to be trusted unless one is able to obtain valid and verifiable supporting evidence.

11 The author once worked with a well-known economist on his economic model—a multiple regression that was run monthly and was intended to predict economic activity. However, the economist would review the results and radically change those with which he did not agree. It seemed that he used the model to display sophistication and credibility, but he always reverted to his old rules of thumb.

12 Wing, J. M.; "Trustworthy AI," *Communications of the ACM*, vol. 64, iss. 10, 2021, p. 64–71

13 *Ibid*.

14 *Op cit* Axelrod 2012

15 *Ibid*.

16 Seshia, S. A.; D. Sadigh; S. Sankar Sastry; "Toward Verified Artificial Intelligence," *Communications of the ACM*, vol. 65, iss. 7, 2022, *https://cacm.acm.org/magazines/2022/7/262079-toward-verified-artificial-intelligence/fulltext*

17 Axelrod, C. W.; *Computer Effectiveness: Managing the Management/Technology Gap*, Information Resources Press, USA, 1979

18 Axelrod, C. W.; "The User View of Computer System Availability," *Journal of Capacity Management*, vol. 2, iss. 4, 1985, p. 340–362, *https://www.researchgate.net/publication/346667073/the_user_view_of_computer_system_availability*

19 Axelrod, C. W.; "The User's View of Computer System Reliability," *Journal of Capacity Management*, vol. 2, iss. 1, 1983, p. 24–41, *https://www.researchgate.net/publication/346667090/the_user's_view_of_computer_system_reliability*

20 *Op cit* Wing

21 Srinivasan, R.; A. Chander; "Biases in AI Systems," *Communications of the ACM*, vol. 64, iss. 8, 2021, p. 44–49, *https://cacm.acm.org/magazines/2021/8/254310-biases-in-ai-systems/abstract*