AI Bookie

At ai.sciencebets.org you can make your own predictions, challenge another prediction and turn it into a bet, or post a bet of your own. We are here to help. So, place your bets at ai.sciencebets.org!

■ *The AI Bookie column documents highlights from AI Bets, an online forum for the creation of adjudicatable predictions and bets about the future of AI. Though it is easy to make a prediction about the future, this forum was created to help researchers craft predictions whose accuracy can be clearly and unambiguously judged when a prediction comes due. The bets will be documented online and regularly in this publication in The AI Bookie. We encourage bets that are rigorously and scientifically argued. We discourage bets that are too general to be evaluated or too specific to an institution or individual. The goal is not to continue to feed the media frenzy and pundit predictions about AI, but rather to curate and promote bets whose outcomes will provide useful feedback to the scientific community.*
Place your bets! Please go to ai.sciencebets.org.

AI Bookie: Will a Self-Authorizing AI-Based System Take Control from a Human Operator?

Donald Sofge, W. F. Lawless, Ranjeev Mittu

Voices are rising about the loss of rigor in AI (for example, see the work of Lipton and Steinhardt [2018]), so the AI Bookie column in *AI Magazine* plans to counter these voices with adversarial views expressed through formal bets. Details are given in an earlier column (Bollacker, Paritosh, and Welty 2018). These bets, we the authors of this new bet hope, will instill a desire to accelerate the science of autonomy and human-machine teams, as well as caution that human lives are at risk not only if the science of AI autonomy advances too quickly, but also if it does not advance at all. From the worldwide threat assessment of the US intelligence community, presented to the Senate Select Committee on Intelligence (Coats 2019, pp. 15-16):

> The global race to develop artificial intelligence (AI) — systems that imitate aspects of human cognition — is likely to accelerate the development of highly capable, application-specific AI systems with national security implications. … AI-enhanced systems are likely to be trusted with increasing levels of autonomy and decision-making, presenting the world with a host of economic, military, ethical, and privacy challenges. Furthermore, interactions between multiple advanced AI systems could lead to unexpected outcomes that increase the risk of economic miscalculation or battlefield surprise.

In the following bet, we hope readers note that we are keenly aware of the risk of surprise posed by what is at stake if we scientists oversell the value of AI while ignoring its dangers, or the greater risk in sitting on the sidelines and not participating in the race described by Coats. At the same time, however, we recognize that at least one of our arguments is likely to be flawed; thus, we also welcome from readers their comments, their clarifications — and their side bets, too.

The Bet

Within 5 years from the publication of this bet, humans will permit AI-enabled systems to self-authorize taking responsibility from their human operator in a noncontrived, nonacademic setting.

Adjudication Criteria

The real issue here is whether the machine can, or will, be permitted to take control against the will of the human operator. Evidence for support of the bet in favor includes recent implementations of AI assistance in commercial systems such as lane assist and emergency braking in automobiles for distracted or errant drivers, and the automatic ground control avoidance system (Auto-GCAS) for regaining control of military aircraft from unconscious fighter pilots. If an autonomous AI system could detect malevolent intent on the part of the human operator, for example, it could take control away from a suicidal and homicidal copilot, such as the copilot who committed suicide and mass murder in 2015 by crashing his Germanwings airliner, killing all aboard. The distinction to be made here is whether within the next 5 years an AI system will be designed to take control against the will of the human. With Auto-GCAS, the assumption is that the human fighter pilot is unconscious, disabled, or otherwise unable to fly the plane safely. Were the pilot able, the assumption is that the pilot would prefer to not crash the plane and avoid loss of his or her own life and possibly the lives of others. In a remarkable scene in *2001: A Space Odyssey,* the deviant computer HAL 9000, when asked to open the pod bay doors, responded, "I'm sorry Dave, I'm afraid I can't do that." We have automatic steering correction built into many new vehicles, but the human can easily counter the motion if desired. In the case of the Germanwings crash, if the AI or autopilot were to take control away from the copilot to save lives, then such a system would address the key requirement: taking control to counter a malicious (or intentionally ignorant) human operator. A single example of an AI taking control from a human operator in a noncontrived, nonacademic setting to save lives will settle the bet in favor of the pro side. The lack of such an example will settle the bet in favor of the con side.

Pro bet: Adjudication criteria accepted.

Con bet: Adjudication criteria accepted.

For: W. F. Lawless

Living electromechanical entities, known as humans, are at the beginning stages of teaming with mobile electromechanical entities, known as machines or robots. Humans, not machines, are the primary cause of accidents. Humans, not machines, get distracted, drowsy, inebriated, angry, suicidal …. In my view, as part of a team, machines are more likely to save rather than threaten human lives (Lawless et al. 2017). But based on the adjudication criteria established by the referee, will we humans allow machines to override a willful human operator intent on harming others? Before answering that question, I review what humans are doing now; afterward, I briefly consider accountability from the consequence of a machine acting against the will of its human-operator teammate.

Automobiles

AI integrated into the electromechanical systems of cars is already helping humans with lane assist, predictive maintenance, insurance claims, and manufacturing. AI is protecting or saving human lives by detecting drowsiness (Novosilska 2018), providing emergency braking, and adjusting speed in construction zones (Krisher 2018). Moreover, based on the National Highway Traffic Safety Administration's news that 10,874 deaths occurred in the United States because of drunk driving in 2017, Volvo is designing cars that limit speed or park in a safe place "to intervene if a clearly intoxicated or distracted driver does not respond to warning signals and is risking an accident involving serious injury or death" (Frangoul 2019).

US Air Force Fighter Planes

Before Auto-GCAS was deployed to save pilot lives, there were dozens of cases ranging from G-induced loss of consciousness to cockpit decompression, hypoxia, and spatial disorientation; since deployment, these systems have saved lives (Lemoine 2009). In the case of a probable ground collision of the new F-35, Auto-GCAS activates, takes control from the pilot, and returns the plane to a safe altitude and attitude until the pilot recovers (Casem 2018).

What about a willful, malicious human operator? In 2015, a Germanwings airliner was flown into the ground by its copilot, who committed suicide and killed all 150 aboard (French Civil Aviation Safety

Investigation Authority 2016). In 2014, Malaysia Airlines flight MH370 manually deviated from its flight path and disappeared with 239 on board (Ministry of Communications and Multimedia 2018). In 2015, a train's engineer allowed his Amtrak Northeast Regional train to speed up in a curve until it derailed in Philadelphia, killing 8 and injuring more than 200 (National Transportation Safety Board 2016). In these and numerous other noncontrived, nonacademic settings, the technology exists for a machine to authorize itself to take limited control when its human-operator teammates are willfully, or ignorantly, threatening human life. Limited control might mean straight and level flight by an airliner, as with Auto-GCAS, until AI control is relinquished; it could mean an AI-controlled landing at the nearest airport for a commercial airliner; or it could mean an AI-controlled emergency train stop.

Human-Machine Teams

In 2018, a pedestrian was killed by an Uber self-driving car (National Transportation Safety Board 2018). The car detected the pedestrian 6 seconds before impact and selected the brakes 1.3 seconds before impact, but Uber engineers had disabled the emergency brakes to improve the car's ride. The car's human operator detected the pedestrian 1 second before impact and hit the brakes 1 second after impact. Clearly, the car performed as designed and was faster than the human operator. However, as part of a human-machine team, the car failed to alert its human teammate to life-threatening danger seconds earlier than the human detected the pedestrian, an easy software and engineering fix. But human-machine teams offer an unexpected opportunity. When machine learning is used by a machine to learn its role as a teammate, implicitly the machine knows what the human is supposed to do (Lawless et al. 2017). With that knowledge, the machine has the tools it needs to know when to intercede partially or fully when its human operator's behavior becomes inappropriate or goes awry.

The counterargument is that a bottleneck arises from the complexity and costs of validating these systems. Not to diminish this very important issue but rather to keep it in perspective, according to the *New York Times* investigation of the 2018 pedestrian death caused by the Uber self-driving car (Wakabayashi 2018), the self-driving cars of Waymo, Uber's competitor, drove an average of nearly 5,600 miles before the driver had to take control from the computer to steer out of trouble. As of March 2018, when the pedestrian was killed, Uber was struggling to meet its target of 13 miles per intervention in Arizona. As incompletely and poorly trained as the Uber car was, it still acted just as it had been designed.

As Coats (2019) noted, the time for autonomous machines is approaching. As well, the time still available for making critical decisions to defend our nation is decreasing. Hypersonic weapons are in development (Magnuson 2019). The US Department of Defense has been modernizing nuclear command, control, and communications systems (US Department of Defense 2018, p. 6), and the first steps are already operational (US Strategic Command Public Affairs 2019). Autonomous swarms of underwater vehicles are in development (Mishra 2019). Retired bipartisan politicians consider that the threat of nuclear war from a mistake is now approaching a "perilous precipice" (Shultz, Perry, and Nunn 2019). As with the Uber self-driving car, because human decision making is slower and more error prone, sooner rather than later machines may need to be designed that can self-authorize to make critical decisions under NC3 procedures but with humans outside of the decision loop (Lawless et al. 2019).

The Montreal declaration for responsible AI demands that AI systems respect human autonomy (University of Montréal 2017):

> AIs must be developed and used while respecting people's autonomy, and with the goal of increasing people's control over their lives and their surroundings. AIs must allow individuals to fulfill their own moral objectives and their conception of a life worth living.

But does this declaration mean that we should build autonomous beings morally superior to humans (McEwan, quoted in Winkler 2019)?

Accountability

Autonomy raises serious issues of accountability (see, for example, the work of Osoba [2019]). In the Uber case, human engineers had disabled the car and made moot the issue of accountability (National Transportation Safety Board 2018). What if, as at present, an AI cannot explain its decisions or actions (Pearl and Mackenzie 2018)? Or, what happens when an AI machine takes control in a noncontrived, nonacademic setting, such as did Hal 9000, the computer that threatened the very humans it was meant to protect, but then the outcome is arguably worse than had it not taken control from its human operator? Who might be held responsible in a court or a court-martial? Current systems of liability accrue to the people, businesses, and industries or militaries that caused the harm. Limited liability legislation may be necessary to encourage businesses to build autonomous machines, similar to the Price-Anderson Nuclear Industries Indemnity Act that allowed development of nuclear power (Insurance Information Institute 2019).

Summary for the Pro Bet

Moral and ethical risks exist with many new technologies (Castelvecchi 2019). Artificial metabolism may add to the threat posed by AI (for example, Hamada et al. 2019). Yet, with the decreasing time available to make critical decisions affecting society and national defense, the question remains: Are we humans ready to permit an AI-based system to self-authorize, taking limited control in a noncontrived, nonacademic

Illustration courtesy James Gary.

setting as when a human operator threatens or fails to protect human lives? I am betting that in the next 5 years, the answer will increasingly be yes: it has already happened on a limited basis, and it is a context evolving before our human — and artificial — eyes.

Against: Ranjeev Mittu

The argument that we will permit an "autonomous machine to self-authorize taking control in a limited context" proposes a state that I do not believe we will be able to fully reach within the next 5 years. I agree with the argument that AI will increasingly self-authorize taking control, but I strongly believe that this will happen only in very limited cases, when the conditions for taking control have been carefully determined, tested, and validated within the AI-based system and known in advance — in other words, when the AI has been carefully programmed a priori. I argue that the cases mentioned in the Pro bet fit within this category. But I do not agree that a case for AI autonomy interesting enough to meet the adjudication criteria for a noncontrived, nonacademic setting will occur in the next 5 years.

To fully reach a state in which AI can autonomously self-authorize taking control from a human operator in a noncontrived, nonacademic setting, the research community will need sufficient time to develop an AI system that can validate its operation and check for singularities that may be affected by poor-quality training data; data distribution biases that may not necessarily reflect the operating environment (Galston 2018); and system complexity, such that employing machine learning is intractable due to the complexity of the interactions that occur and that are so poorly understood that the data collection to properly train the AI to achieve full autonomy over the next 5 years is likely to remain infeasible (Hintze 2017).

Maritime Domain Awareness

Commercial shipping vessel movements around the world are tracked using various techniques such as the automated information system (AIS). The AIS data contain information pertaining to a vessel such as coordinates, course, and speed (International Maritime Organization 2019). The research community has explored the application of machine learning to learn patterns in vessel traffic to identify ships involved in illicit activities. It proved to be a difficult problem to model using machine learning (Newman 2019). One of the problems uncovered in applying machine

learning was that the models were developed from poor data. As another example, AIS can be knowingly spoofed. Furthermore, one could imagine an adversarial attack on AIS data (for example, Kessler, Craiger, and Has 2018). In this case, AI should not be allowed to authorize itself to make a decision even after concluding that a vessel was a potential threat based on apparent anomalies in its kinematic patterns that could have been planted by an adversary.

Distribution Biases

As of now, facial recognition technology works better for men than women and for people of lighter complexion than for those of color (Galston, 2018). The danger is that false positives infected with systematic biases must not be ignored. Giving AI-enabled systems too much authority too soon is an overreach that may impair the rights of citizens. Technology companies fear a backlash from overreaching; that is one reason that they are developing codes of ethical behavior. An uneven application of these codes, however, offers a role for new laws and for government oversight.

System Complexity

HAL 9000 was devised by the science fiction author Arthur C. Clarke (Hintze 2017) and later brought to the screen by movie director Stanley Kubrick in *2001: A Space Odyssey*; this fictional computer provided a great example of a system that takes control from a human but fails to protect human life because of unintended consequences (Hintze 2017). In many complex systems — the *RMS Titanic,* NASA's space shuttle, the Chernobyl nuclear power plant, the two Boeing 737 Max planes — engineers stack together one layer after another of various components. In these and other cases, the engineers may have known how each aspect of a system worked individually, but they did not know well enough how all the subsystems worked together. The result was complex systems that could never be fully understood and that even could fail in unexpected ways. In each of these and many other tragedies — a sunken ship, two shuttles lost, radioactive contamination spread across Europe and Asia, two planes falling from the sky — a set of relatively small system failures combined to create a tragedy.

Summary for the Con Bet

When human lives are at stake, it would be nice to have a system that rescues and safeguards them until authorities could take over, a future dream. But the biases in machine learning software, their common lack of quality data, and the steps that an autonomous system may take unexpectedly and spontaneously likely preclude our human designers from designing a machine to self-authorize taking responsibility from its human operator in a noncontrived, nonacademic setting over the next 5 years.

References

Bollacker, K.; Paritosh, P.; and Welty, C. 2018. The AI Bookie. Place Your Bets: Adversarial Collaboration for Scientific Advancement. *AI Magazine* 39(4): 84–7. doi.org/10.1609/aimag.v39i4.2837

French Civil Aviation Safety Investigation Authority. 2016. Final Report. Accident on March 24, 2015 at Prads-Haute-Bléone (Alpes-de-Haute-Provence, France) to the Airbus A320-211 Registered D-AIPX Operated by Germanwings. Paris: Government of France.

Casem, G. 2018. F-35s Begin Auto GCAS Test Flights. Washington, DC: US Air Force Public Affairs.

Castelvecchi, D. 2019. AI Pioneer: The Dangers of Abuse Are Very Real. *Nature News Q&A* (April 4). doi.org/10.1038/d41586-019-00505-2.

Coats, D. R. 2019. Statement for the Record: Worldwide Threat Assessment of the US Intelligence Community. Washington, DC: US Senate Select Committee on Intelligence.

Frangoul, A. 2019. Volvo to Put Cameras and Sensors in *Its Cars to Tackle Drunk Driving. CNBC (March 21)*. www.cnbc.com/2019/03/21/volvo-to-put-cameras-and-sensors-in-its-cars-to-tackle-drunk-driving.html

Galston, W. A. 2018. *Why the Government Must Help Shape the Future of AI*. Washington, DC: Brookings Institute.

Hamada, S.; Yancey, K. G.; Pardo, Y.; Gan, M.; Vanatta, M.; An, D.; Hu, Y.; Derrien, T L.; Ruiz, R.; Liu, P., Sabin, J.; and Luo, D. 2019. Dynamic DNA Material with Emergent Locomotion Behavior Powered by Artificial Metabolism. *Science Robotics* 4(29): eaaw3512.

Hintze, A. 2017 What an Artificial Intelligence Researcher Fears about AI. *The Conversation* (July 13). theconversation.com/what-an-artificial-intelligence-researcher-fears-about-ai-78655.

Insurance Information Institute. 2019. *Insurance Coverage for Nuclear Accidents*. New York: Insurance Information Institute.

International Maritime Organization. 2019, *AIS Transponders*. London, UK: International Maritime Organization.

Kessler, G. C.; Craiger, J. P.; and Haass, J. C. 2018. A Taxonomy Framework for Maritime Cybersecurity: A Demonstration Using the Automatic Identification System, TransNav. *International Journal on Marine Navigation and Safety of Sea Transportation* 12(3). doi.org/10.12716/1001.12.03.01

Krisher, T. 2018. New Cars Are Quickly Getting Self-Driving Safety Features. Phys.org (March 27). phys.org/news/2018-03-cars-quickly-self-driving-safety-features.html.

Lawless, W. F.; Mittu, R.; Sofge, D.; and Russell, S., editors. 2017. *Autonomy and Artificial Intelligence: A Threat or Savior?* New York: Springer. doi.org/10.1007/978-3-319-59719-5

Lawless, W. F.; Mittu, R.; Sofge, D. A.; and Hiatt, L. 2019. Artificial Intelligence, Autonomy, and Human-Machine Teams: Interdependence, Context, and Explainable AI. *AI Magazine* 40(3).

Lemoine, C. W. 2009. What Exactly Is Auto GCAS? *Fighter Sweep*. fightersweep.com/3955/exactly-auto-gcas/.

Lipton, Z. C., and Steinhardt, J. 2019. Troubling trends in machine learning scholarship. *ACM Queue*.

Magnuson, S. 2019. Hypersonic Jet Project Reaches Major Milestone. *National Defense* (April 11). www.nationaldefensemagazine.org/Articles/2019/4/11/Hypersonic%20Jet%20Project%20Reaches%20Major%20Milestone.

Mishra, S. 2019. Could Unmanned Underwater Vehicles Undermine Nuclear Deterrence? *The Strategist* (May 8). www.aspistrategist.org.au/could-unmanned-underwater-vehicles-undermine-nuclear-deterrence/.

First Call for Nominations for the 2020 AAAI Executive Council Election

The 2020 Nominating Committee is seeking nominations from the AAAI membership for the positions of AAAI President-Elect and Executive Councilor. In 2020, AAAI members will elect one individual to serve a two-year term as president-elect, followed by two years as president, and finally, two years as immediate past president. In addition, members will elect four new councilors to serve three-year terms on the AAAI Executive Council. All elected officers and councilors are required to attend all council meetings each year (usually 1–2 in person and 2–3 via telecon), and actively participate in AAAI activities. Nominees must be current members of AAAI. The Nominating Committee encourages all regular AAAI members in good standing to place an individual's name before them for consideration. (Student and institutional members are not eligible to submit candidates' names.) The Nominating Committee, in turn, will nominate two candidates for president-elect and eight candidates for councilor in early spring. In addition to members' recommendations, the committee will actively recruit individuals in order to provide a balanced slate of candidates. AAAI regular members will vote in late spring, and the new members of the Executive Council will be installed in the summer of 2020.

To submit a candidate's name for consideration, please send the following information to Carol Hamilton, Executive Director, AAAI, 2275 East Bayshore Road, Suite 160, Palo Alto, CA 94303; by fax to 650/321-4457; or by email to hamilton@aaai.org:

Name
Affiliation
City, State or Province, Country
Email address
URL
Year of membership in AAAI
Approximate number of AAAI publications
At least two sentences describing the candidate and why he or she would be a good candidate

Please include any additional information or recommendations that would be helpful to the Nominating Committee. Nominators should contact candidates prior to submitting their names to verify that they are willing to serve, should they be elected. The deadline for nominations is March 1, 2020.

Ministry of Communications and Multimedia. 2018. *MH370 Safety Investigation Report*. Kuala Lumpur: Government of Malaysia.

National Transportation Safety Board. 2016. Derailment of Amtrak Passenger Train 188. www.ntsb.gov/investigations/AccidentReports/Pages/RAR1602.aspx.

National Transportation Safety Board. 2018. Preliminary Report Released for Crash Involving Pedestrian, Uber Technologies, Inc., Test Vehicle. www.ntsb.gov/news/press-releases/Pages/NR20180524.aspx.

Newman, N. 2019. Cyber Pirates Terrorising the High Seas. *E&T Engineering and Technology* (April 18). eandt.theiet.org/content/articles/2019/04/cyber-pirates-terrorising-the-high-seas/.

Novosilska, L. 2018. 5 Ways Artificial Intelligence is Impacting the Automotive Industry. *Ignite* (November 30). igniteoutsourcing.com/automotive/artificial-intelligence-in-automotive-industry/.

Osoba, O. A. 2019. *Keeping Artificial Intelligence Accountable to Humans*. Santa Monica, CA: RAND Corporation.

Pearl, J., and Mackenzie, D. 2018. AI Can't Reason Why. *Wall Street Journal* (May 18).

Shultz, G. P.; Perry, W. J.; and Nunn, S. 2019. The Threat of Nuclear War Is Still with Us. *Wall Street Journal* (April 10).

University of Montréal. 2017. *Montréal Declaration for Responsible Development of Artificial Intelligence*. Montréal: University of Montréal.

US Department of Defense. 2018. *Nuclear Posture Review*. Washington, DC: Office of the Secretary of Defense.

US Strategic Command Public Affairs. 2019. USSTRATCOM Announces Initial Operational Capability of NC3 Enterprise Center. www.stratcom.mil/Media/News/News-Article-View/Article/1805006/usstratcom-announces-initial-operational-capability-of-nc3-enterprise-center.

Wakabayashi, D. 2018. Uber's self-driving cars were struggling before Arizona crash. *New York Times* (March 23).

Winkler, E. 2019. Novelist Ian McEwan Ponders Our AI Future. *Wall Street Journal* (April 11).

William Lawless blew the whistle on Department of Energy (DOE) mismanagement of military radioactive wastes. After his PhD, he joined DOE's citizen advisory board at its Savannah River Site where he coauthored over 100 recommendations on its cleanup. His research today is on interdependence for teams (human-machine teams). He is also a professor at Paine College.

Ranjeev Mittu is the Branch Head for the Information Management and Decision Architectures Branch within the Information Technology Division, US Naval Research Laboratory. His research expertise is in multi-agent systems, artificial intelligence, machine learning, data mining, pattern recognition and anomaly detection.

Donald Sofge is a computer scientist and roboticist at the U.S. Naval Research Laboratory (NRL) researching teams and swarms of autonomous robotic systems.