Ethics Education of Human Factors Engineers for Responsible AI Development

Esa M. Rantanen[1], John D. Lee[2], Katherine Darveau[3][4], Dave B. Miller[5],
James Intriligator[4], and and Ben D. Sawyer[5]
[1]Rochester Institute of Technology, [2]University of Wisconsin–Madison, [3]GE Aviation,
[4]Tufts University, [5]University of Central Florida

This panel discussion is third in a series examining the educational challenges facing future human factors and ergonomics professionals. The past two panels have focused on training of technical skills in data science, machine learning, and artificial intelligence to human factors students. This panel discussion expands on these topics and argues for a need of new and broader training curricula that include ethics for responsible development of AI-based systems that will touch lives of everybody and have widespread societal impacts.

INTRODUCTION

At the past two Human Factors and Ergonomics Society (HFES) Annual Meetings we have presented panel discussions on the educational needs of the future human factors and ergonomics (HF/E) workforce, focusing on educational gaps in the areas of data science, machine learning (ML), and artificial intelligence (AI) (Hannon et al., 2020, 2019). It appears that these panels have provided some impetus to re-examine the education of HF/E professionals more broadly as well (Zhang & Chignell, in press). However, as typically is the case in HF/E, technological advances far outpace the ability of the discipline to provide essential input to human-centered development of systems. This is particularly noticeable, noteworthy, and indeed critical with AI.

In our past panels we have contrasted the “1st generation” (cockpit) automation with AI/ML-based "2nd generation" automation, focusing on the opaqueness of the latter and the challenges to human-automation interaction (HAI) it poses. This focus may be too narrow. The biggest problem may not be in the “box” after all, whether it is black or any other color, but that of scale. If old forms of automation were singular devices (e.g., an autopilot) or vehicles (e.g., a Mars rover), new automation is akin to a “Humongous Fungus” (*armillaria ostoyae*), covering a huge “area” in terms of impacted people and their myriad activities with networked devices and vast amounts of user data collected. Hence, the human factors challenge is less that of HAI in different tasks, but more about the role of humans in the evolving and interconnected technological society in general.

The HFES mission statement already is human-centered: designing for people, and designing systems that work for humans. HFES has members who focus on micro-cognition and the details of how people extract information from displays. Other members focus on macro-cognition and sociotechncial systems, where the focus is often on teams and organizations. A particular challenge with ubiquitous AI is that it exerts influence simultaneously at the micro *and* macro levels. This panel will extend the discussion to the direction of what is the ultimate human purpose. Is it just that of a consumer, something to be freed of all effort, and perhaps even all initiative? Or should technology be designed to support continuous human development? What do these challenges mean for the development of HF/E curricula and the education of future HF/E professionals?

PANELISTS’ POSITIONS

Katherine Darveau, GE Aviation and Tufts University. The intersection of my experiences in industry and academia show several potential benefits to integrating AI, Human Factors, and domain expertise. Collaboration between industry and academia provides a mutual benefit to students, universities, and employers. Sharing of data, data collection methodologies, and reporting interfaces gives HF/E students the opportunity to set up an AI problem in the context of a real-world scenario, that may contain a broad array of HF/E challenges that are not well understood by the industry. A critical benefit of HF/E integration into AI is investigating and understanding the current state limitations of data collection and implications for a potentially automated system.

While standard human-in-the-loop systems require human intervention or oversight to prevent improper automated decision making, less focus is placed on the role of humans (e.g., domain subject matter experts) in the design of the system and modeling approaches. Inherent to a HF/E curricula are human-centered, investigative, and causal approaches to problem solving, which may be lacking in the majority of traditional AI applications. Therefore, there is a clear benefit to a collaboration among HF/E professionals, domain experts, and data scientists to improve AI capability and find the right balance of human-machine teaming that addresses meets both performance and ethics goals. It may be worthwhile to demonstrate the effectiveness of this collaboration at the college level, with industry partnerships and cross-departmental projects.

My dissertation seeks to find the right balance of human-machine teaming for safety event classification that achieves desired performance. In so many examples of AI with human oversight, there are trade-offs between level of human oversight and performance that need to be evaluated based on outcome goals, available resources, and other business-related factors. The key is to iteratively test and improve our decision tools to determine where and why AI may not be as successful as a human. In some cases, automation can not only expedite the process but may outperform humans due to the nature of the task, whether mundane/monotonous or of high cognitive load. In high-risk industries, the “right” balance should heavily prioritize the probability and severity associated with failure. An example from the medical industry is the use of clinical diagnostic support system (CDSSs) that help to diagnose a patient’s illness and/or assist in making decisions about treatment plans. In many cases, CDSSs are used to make an initial recommendation but the clinician proves a final evaluation and decision (Anooj, 2012; Castaneda et al., 2015; Hasan & Padman, 2006; Scheitel et al., 2017; Walczak, 2005; Walczak, Pofahl, & Scorpio, 2003). In the aviation industry, human-machine teaming for operational and safety event classification has proven more successful than automated methods alone (Bluvband & Porotsky, 2012; Labarga & Friedman, 2018).

A major limitation of the success of AI is the quality of the source data, which can vary significantly if entered by humans. While some data quality issues can be counteracted by the development of an effective AI system or incorporation of human intervention, these approaches may not guarantee a perfectly successful system. It is beneficial for both system and user (person who enters the data) to engage in continuous learning. The concept of "usable" or user-interactive machine learning yields continuous development for both the human and the automated system. In a traditional model, iterative system improvements require someone well-versed in machine learning to interpret end user behavior and tweak model parameters. A more desirable end state is to eliminate the need for a ML expert and allow the system to directly learn from the user, and vice versa. The concept of usable machine learning provides an interactive process that allows user to clarify or correct data inputs, provide additional detail as requested, and/or approve or deny an AI prediction. In addition to influencing user behavior to yield better data quality, the user's domain knowledge is also improving the predictive capability of the system (Amershi, Cakmak, Knox, & Kulesza, 2014).

James Intriligator, Tufts University. Henry Ford famously quipped "Any customer can have a car painted any color that he wants—so long as it's black." Today, we recount the Ford tale and laugh about the early days of design and the lack of customer/user/human focus. Happily, we now have technologies and infrastructures that let us cater to customer desires. We can research and understand the specific desires of our customers (or users) and design to meet their desires. In HF/E this approach is referred to as "human-centered design" (HCD).

We teach our students to put the user at the center of design: to study them, empathize with them, interview them, understand them and their roles, tasks, relationships and journeys. And, from this design research, we create a detailed user persona. This persona then guides all subsequent design. We know that this approach works well, but HCD also has potential dangers: What if we get the wrong persona? What if we leave out or exclude another important persona? What if our design inadvertently makes life worse for another persona?

Ending up with the wrong color car is not great but narrowly focused HCD can lead to far more serious errors: loan decision systems with racial biases or face recognition systems (trained only Caucasian faces) that incorrectly recognize non-white individuals. Often such failures (found in numerous products, systems, or processes) began at the earliest stages of research and design. As we move into a future of mass-customization, mass-use, and digital artifacts, such problems will compound and appear more frequently and with more devastating impacts.

We can help mitigate such design failures by using the "persona multiplication method" (PMM). Begin with standard design research to create a human-centered user persona but do not stop there. Instead, pause, reflect, and examine the persona with an eye towards multiplying it. Look at the persona and ask, "What if I varied some of the dimensions—the gender, age, race, or class?" and "What personae will be excluded if we proceed with only this persona?" and "What disadvantaged or marginalized groups would be excluded, harmed, or further disadvantaged, if we proceed with only this one persona?" In many cases, we might decide to include multiple (or wider) personae in our design process.

PMM helps identify and avoid the dangers of narrow HCD and we can expand our persona sets to include other personae that might have been missed, excluded, or disadvantaged. For example, imagine designing an ML-based handwriting recognition system. As we consider gathering training data, we might realize that our persona was (implicitly or explicitly) a right-handed individual. Using only this persona will exclude millions of left-handers. Thus the persona needs to be multiplied (or at least expanded) to avoid excluding important groups. The same approach should be applied across a wide range of important spaces. We might design different user interfaces that will not leave behind the visually impaired, the non-native speakers, or older individuals. In the realm of machine learning we will begin by choosing training-sets that span the wider range of personae who might be using (or processed by) the system.

HCD has served designers well, but, as products and systems become more customizable and more widely used, we need to move away from the narrow persona that HCD advocates. We must examine the multidimensional realm of possible personae and ask whether our persona/design might be leaving people behind by inadvertently creating an advantage for some and excluding or disadvantaging others. This step will allow us to redirect our design efforts to create more ethical products, systems, and processes.

John D. Lee, University of Wisconsin–Madison. The provocatively titled book, "Weapons of Math Destruction," describes damage machine learning algorithms can do when deployed at scale (O'Neil, 2016). Their destructive capacity stems from their scale, opaqueness, and potential for societal damage. One notorious example is the COMPAS system. It uses information about convicted criminals to predict recidivism and guide parole decisions (Dressel & Farid, 2018). Such systems can profoundly affect the lives of millions of people. Because these systems often suffer from racial and gender biases, they can inflict societal-scale damage.

The scale considered in "Weapons of Math Destruction" is the *breadth* of influence, defined by the number of people affected. Another dimension of scale concerns the *depth* of influence. Here depth is how deeply the algorithms affect people's lives, attitudes, and thinking. One avenue for deep influence of machine learning is through smartphones. Smartphones have become ubiquitous with 95% of teens report having access and 45% using constantly (Firth et al., 2019). Smartphones are used frequently: an average of 233 minutes per day, with 86 pickups, and 90 notifications (Ellis, Davidson, Shaw, & Geyer, 2019). These interactions can produce acute and chronic changes in cognition, such as diminished ability for sustained concentration, and even changes brain structure (Firth et al., 2019).

Algorithms embedded in smartphones deepen this influence by selecting news and mediating social interactions: 62% of US adults received news through social media (Carlson, 2018). Algorithms presenting this news are tuned to engage rather than educate people and one consequence is polarization and the spread of misinformation (Del Vicario et al., 2016). These algorithms also personalize advertising to incite purchasing, sometimes to excess. One explanation for this is that reinforcement learning algorithms can maximize their cost function by either meeting people's needs or by changing those needs (Russell, 2019). Increasing polarization and excessive purchasing may reflect algorithms guiding people to serve the algorithm, rather than the algorithms serving people. The power of algorithms to create deeply personalized interactions can either protect us from our weakness or prey on them (Zuboff, 2019).

What can be done? Human factors practitioners can become familiar with algorithmic bias and methods to assess fairness (Albarghouthi & Vinitsky, 2019). No single fairness metric addresses algorithmic bias: their application requires understanding the context of use, which human factors practitioners are well-positioned to address. Human factors practitioners also understand the principles of cognitive psychology, choice architecture, and persuasive computing that deepen algorithm influence (Eyal, 2014; Thaler & Sunstein, 2008). As advocates for people interacting with AI at a broad and deep scale, we can guide AI design to serve people rather than treat their behavior as raw material to be mined and manipulated (Zuboff, 2019).

Dave B. Miller, University of Central Florida. The discipline of human factors is dedicated to human safety, security, and wellbeing, applying often hard-won knowledge to the human-technology interface. With sophisticated AI systems shaping our beliefs and desires, helping us drive our cars, and policing us (Benjamin, 2019), ethics training must be reimagined to prepare students to face new challenges presented by these new technologies that can act with independent agency (Miller, 2016) and which can reproduce value choices and biases at societal scale. AI systems can be both persuasive (Fogg, 2003) and coercive (Eubanks, 2018), and as part of the fabric of modern society, there is no avoiding them.

Automated systems provide an ability to exert force at a temporal and spatial distance: decisions made in a cubicle today are performed as actions in some other place, at some future time. AI systems, which increase the degree of remove from human control compared to less flexible forms of automation, demand even greater scrutiny as they can even more easily exhibit unpredictable behavior. Considering how the selection of training sets and other features of AI development almost inevitably encode the heuristics and biases of system creators, the parameters of organizational culture are even more pertinent issues in terms of the risks cultural deficiencies pose.

To this end I propose expanding the scope of engineering and design ethics to beyond merely avoiding "defective products" to more universally include a focus on technical culture, and training students about how to create positive technical cultures and avoid or change negative ones. The effect of organizational culture on safety has been the subject of significant investigation in many areas, and especially in aerospace where Diane Vaughan coined the term "normalization of deviance" (Vaughan, 1996). This term describes the process in which originally unacceptable behavior becomes routine—often with disastrous results such as accepting significant risks or allowing callous indifference to biases to persist. With AI systems that make decisions and take action in response to complex programs or which are trained and therefore exhibit a degree of autonomy and unpredictability, the organizational culture almost inevitably will influence the behavior of the system, and thus issues of culture are now more important than ever given the spread of AI systems into every corner of our lives. Analyzing the crashes of Boeing's 737-MAX aircraft, the killing of a pedestrian by a prototype Uber autonomous vehicle, and even the race-biased differences in performance of image processing algorithms shows that faults in the culture of technology companies translate into faulty products of design.

Engineering and design ethics can be further extended from the regime of avoiding the negative to pursuing the positive. As Willy Wonka said: "we are the dreamers of dreams"—as creators of the things that make up our built environment, and now touch almost every aspect of modern life, it is not too far a reach to consider pro-social and pro-environmental aims as integral to the human factors mission. Considering too small a scope leaves us open to making life easier in some ways while more difficult in others; or better for some and worse for others. Given the power of AI systems to shape our information environment and increasingly our physical surroundings, we need to take care in employing such systems when they can have outsize adverse effects on the already marginalized, and design these systems with respect to ensuring they actively do good, remedying extant disparities rather then exacerbating them and widening the chasms between the haves and have-nots, the favored and the disfavored.

HF/E practitioners, being involved in the design and validation of many products, services, and systems, bear perhaps special responsibility for considering not just the immediate safety risks and threats to usability, but societal risks as well, as they can bring to the fore issues beyond pure engineering concerns. Educating HF/E students to understand the role AI has and will continue to play in shaping society can help to address hazards such as the extension of sexism, race and class bias, wealth inequality, and political polarization. Students, especially engineers, need to be part of scholarly discussions regarding major social problems and how technology can exacerbate or help to remedy them. They cannot graduate retaining the mindset that they are mere technologists and remedying social problems is beyond their purview. Integrating ethics, with a broad social focus, into the curriculum is a way we, as educators, can further the essential work of making the world a safer and more pleasant place, for everyone.

Ben D. Sawyer, University of Central Florida. Our discipline is quite convinced that humans are not difficult to manipulate. We can therefore agree that great caution, and altruistic intent, must prevail as we invest previously unseen amounts of human intellectual capital in the creation of agents capable of manipulating human behavior. Such technologies may undoubtedly serve noble purposes, but their very existence should give us pause. In my recent book chapter on automation, autonomy, and artificial intelligence (A^3), I argue that it is crucial that we design A^3 we will *all* like (Sawyer, Miller, Canham, & Karwowski, 2021).

I also believe we are going to be incapable of not loving forthcoming A^3; our own discipline is currently engaged in helping these systems to be very capable of pushing humanity's many, shockingly exposed, buttons. Loving something you do not like is a special type of hell which I will here suggest that we work hard to not build for ourselves, and those that come after us. Consider: distrust and mistrust of current and previous generation technologies costs lives (Parasuraman & Riley, 1997). A similar catastrophic failure in coming generations of A^3 might negatively alter the very fabric of human experience, restructuring the way we work, live, and even think. My contribution to the panel is a call for change. First, we must retain relevance through aggressive action in three parts: research, application, and education. We must perform research that asks the hard questions. There are a few other engineering-adjacent communities interested in the roots of trust , ethics, success, failure, pleasure, and misery. Here, we dare not simply wave our hands, we must instead use them to excavate the truth and hold it out for all to see. Researchers must arm our practice-oriented colleagues with these truths, and send them into the field to apply them where they can make real impact on the diverse futures available. Human factors has always relied upon its practitioners for the hard work of making the real changes in the world, and we will rely on them again. Finally, we must attract and train those who are not our equals, but our superiors in this task. As directly addressed in previous iterations of this panel, human factors presently runs the risk of losing our ability to guide the development of technology through simple lack of technical ability. This must not happen, and must be counted aggressively in curricula, culture, and through radical inclusivity. Technically adept students, from diverse socioeconomic backgrounds and from every community on earth, must find programs of human factors to be welcoming, challenging, and springboards into careers that effect the change we need.

We must all find human factors to be springboards into careers that are building a better future. Regrettably, cognitive engineering can be used to look for exploits in cognition, human performance does not necessarily need to focus upon enhancement, and the deficit foundations of the psychological portion of our training leaves us well-equipped to find intersections of those deficits with desirable, often profitable, patterns of behavior (Canham & Sawyer, 2020; Sawyer et al., 2021). But our community is also a bastion of careful, ethical consideration. Our "safety science" is well-placed for, and currently engaged in, creating the design decisions, philosophies, and interlock technologies that will become the building blocks of good technological teammates. As a science, we understand that technology alone is not the answer, and that optimal performance comes from hybrid teams: part human, part machine. We understand the dysfunctions of such systems, and the ways to nudge them toward stability and comfort. We understand and can warn of particularly dark futures, including the one in which we all monitor increasingly capable machines for increasingly rare and catastrophic failures (Sawyer & Hancock, 2018), alluded to in practically any applied vigilance paper. We must, however, do even more. I would like to suggest that the discipline consider not merely the outcomes and ethics of the systems we build, but the ethical orientation of our own community

As humans embark on building intelligent teammates, I suggest we turn our community toward pushing for these agents to be designed according to an example we as a community set. Indeed, I argue the discipline might do very well to establish an oath in this regard, a *Primum non Nocere* of technology design. I believe human factors researchers and practitioners would find "first, do no harm" an excellent mantra by which to avoid the dark patterns our training can be turned toward. We already accept this responsibility, engaging in good faith with IRBs, conducting research ethically, and advocating for all of the humans within the systems we design. So many of us have been the rare voice urging consideration of this human factor on projects, pulling or pushing toward those brighter futures. I personally have found that by being the person to speak up, I find allies in my fellow humans, who have some inherent interest. Our community can fortify one another in this common scenario, for there is strength in common resolve. This is the reason for the power in a physician's suggesting that an action might violate her oath. This can be a path for us, as a community, to define and accept great responsibility.

Should human factors as a community work to differentiate itself by embracing a responsibility in navigating humanity toward brighter futures? I argue yes, and further that there is profit and glory in this path; respect for this kind of community commitment has launched other disciplines to new places of regard and utility in society. The details, of course, are crucial. What should such an oath look like, focused upon A^3 and beyond? How do we get our students on board? Our peers? Our teachers? What about *everyone* else?

Esa M. Rantanen, Rochester Institute of Technology. As AI systems become increasingly ubiquitous and capable, and their widespread applications in myriad domains and across all societal elements, the influence of AI, including its unintended and unforeseeable consequences, will also be felt on a larger, societal, scale than has been a case with earlier technologies. The most difficult research questions therefore pertain to scale. Current and future ML/AI-based automation will penetrate every aspect of people's lives, through networked devices (Internet of Things, IoT) and collection of vast amounts data through them, and sharing these data in myriad forms across the IoT and different government and private entities. Past approaches to human-system integration and cognitive systems engineering must be correspondingly "scaled up". As AI will touch the lives of everybody, increasing human variability across heterogeneous user groups presents additional challenges.

Effective communication between AI and human collaborators is critical, including models of the agents' (both human and machine) models of their environment, goals, constraints, and intents, and the moment-to-moment changes in these. Unique human characteristics, most importantly their intuitive decision-making and social- and emotional intelligence, must be represented in a way that is understandable to the machine. What kinds of models of humans do AI systems trained by Imitation Learning or Inverse Reinforcement Learning (Alexander, 2018; Piot, Geist, & Pietquin, 2016) build? How are such models represented for evaluation and validation?

Ubiquitous interactions between two fundamentally different agents, humans as analog beings (Norman, 1998) and digital computers, each relying on imperfect and different but interactively and dynamically shaped models of each other presents many fundamental research problems. The danger is to mistake formal and often numerical models for reality, and build systems based on such deficient specifications of what is real (Christian, 2020). Critical and in-depth study of models, including development of appropriate methods for their study, is therefore paramount for both future development of AI-based technologies and training of those developing them, for "All models are wrong, but some are useful" (Box, 1979).

A particularly difficult challenge for human-AI interaction research is that the problem is multiple-dynamic (Reason, 1990). AI is trained by experience with human interactions, but these interactions are also influenced by the human experience with the AI agent. In terms of models, AI systems model humans as walking neural nets being trained by the systems themselves (Christian, 2020). Development of "supermodels", or models of models, of human-AI interactions is therefore a critical challenge for the HF/E community to tackle. Training of those who will do research with such "supermodels" is of course the first task.

REFERENCES

Albarghouthi, A., & Vinitsky, S. (2019). Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 211–219).

Alexander, J. (2018). Learning from humans: what is inverse reinforcement learning? *The Gradient*. https://thegradient.pub/learning-from-humans-what-is-inverse-reinforcement-learning/.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, *35*(4), 105–120.

Anooj, P. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, *24*(1), 27–40.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new jim code. *Social Forces*, *98*(4), 1–3.

Bluvband, Z., & Porotsky, S. (2012). Advanced text mining algorithms for aerospace anomaly identification. *Advances in Safety, Reliability and Risk Management*.

Box, G. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press.

Canham, M., & Sawyer, B. D. (2020). Human brain electro-optical signals as masint. *American Intelligence Journal*, *125*(4), 40-47.

Carlson, M. (2018, January). Facebook in the News. *Digital Journalism*, *6*(1), 4–20. doi: 10.1080/21670811.2017.1298044

Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., … Suh, K. S. (2015). Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of Clinical Bioinformatics*, *5*(1), 1–16.

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on Facebook. *Scientific Reports*, *6*(1), 37825.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), 1–6. doi: 10.1126/sciadv.aao5580

Ellis, D. A., Davidson, B. I., Shaw, H., & Geyer, K. (2019, October). Do smartphone usage scales predict behavior? *International Journal of Human-Computer Studies*, *130*, 86–92. doi: 10.1016/j.ijhcs.2019.05.004

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Eyal, N. (2014). *Hooked: How to build habit-forming products*. Penguin.

Firth, J., Torous, J., Stubbs, B., Firth, J. A., Steiner, G. Z., Smith, L., … Sarris, J. (2019, June). The "online brain": How the Internet may be changing our cognition. *World Psychiatry*, *18*(2), 119–129. doi: 10.1002/wps.20617

Fogg, B. J. (2003). *Persuasive technology: using computers to change what we think and do*. Morgan Kaufmann.

Hannon, D., Rantanen, E., Sawyer, B., Hughes, A., Darveau, K., O'Donnell, R., … Lee, J. D. (2020). The education of the human factors engineer in the age of data science. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 64, pp. 480–484).

Hannon, D., Rantanen, E., Sawyer, B., Ptucha, R., Hughes, A., Darveau, K., & Lee, J. D. (2019). A human factors engineering education perspective on data science, machine learning and automation. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 63, pp. 488–492).

Hasan, S., & Padman, R. (2006). Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach. In *AMIA Annual Symposium Proceedings* (Vol. 2006, p. 324).

Labarga, J., & Friedman, S. (2018). Automated labeling of helicopter maintenance records with text classification. In *Proceedings of the 74th Annual Forum and Technology Display.*

Miller, D. (2016). AgentSmith: Exploring Agentic Systems. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 234–238).

Norman, D. (1998). *The invisible computer*. The MIT Press.

O'Neil, K. (2016). *Weapons of Math Destruction: How big data increases inequality and threatens democracy*. New York: Crown.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Piot, B., Geist, M., & Pietquin, O. (2016). Bridging the gap between imitation learning and inverse reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(8), 1814–1826.

Reason, J. (1990). *Human error*. Cambridge University Press.

Russell, S. (2019). *Human Compatible: AI and the problem of control*. Penguin.

Sawyer, B. D., & Hancock, P. A. (2018). Hacking the human: the prevalence paradox in cybersecurity. *Human Factors*, *60*(5), 597–609.

Sawyer, B. D., Miller, D. M., Canham, M., & Karwowski, W. (2021). Human Factors and Ergonomics in Design of A^3: Automation, Autonomy, and Artificial Intelligence. In *The handbook of human factors.* Wiley.

Scheitel, M. R., Kessler, M. E., Shellum, J. L., Peters, S. G., Milliner, D. S., Liu, H., … others (2017). Effect of a novel clinical decision support tool on the efficiency and accuracy of treatment recommendations for cholesterol management. *Applied Clinical Informatics*, *8*(1), 124.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New York: Penguin Books.

Vaughan, D. (1996). *The challenger launch decision: Risky technology, culture, and deviance at nasa*. University of Chicago press.

Walczak, S. (2005). Artificial neural network medical decision support tool: predicting transfusion requirements of ER patients. *IEEE Transactions on Information Technology in Biomedicine*, *9*(3), 468–474.

Walczak, S., Pofahl, W. E., & Scorpio, R. J. (2003). A decision support tool for allocating hospital bed resources and determining required acuity of care. *Decision Support Systems*, *34*(4), 445–456.

Zhang, Y., & Chignell, M. (in press). The gap between human factors engineering education and industry needs. In *Proceedings of the 21st Triennial Congress of the International Ergonomics Association.*

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.