Handbook of Human Factors and Ergonomics, Fifth Edition. Gavriel Salvendy and Waldemar Karwowski.

CHAPTER 52

HUMAN FACTORS AND ERGONOMICS IN DESIGN OF A^3: AUTOMATION, AUTONOMY, AND ARTIFICIAL INTELLIGENCE

Ben D. Sawyer, Dave B. Miller, Matthew Canham, and Waldemar Karwowski
University of Central Florida
Orlando, Florida

1 INTRODUCTION

Automation, autonomy, and artificial intelligence (AI) are technologies which serve as extensions of human ability, contributing self-produced, non-human effort (see Figure 1). These three terms encompass a set of computational tools that can learn from data, systems that act in a reasonable, and even human-like manner (Bolton, Machová, Kovacova, & Valaskova, 2018; Dash, McMurtrey, Rebman, & Kar, 2019; Shekhar, 2019). Computing of this nature has been pursued at least since the 1950s, when Simon predicted machines "capable … of doing any work a man can do" (Chase & Simon, 1973), and today such envisioned technology appears under the moniker Artificial General Intelligence (AGI). The desire for synthetic intelligent creations has been a staple of human desire for much longer, in various forms (Hancock et al., 2011; Schaefer et al., 2015). While AGI remains, at present, just a dream. A number of promising, and promised, future technologies under development require machines to learn, understand, and adapt to novel situations with at least the flexibility humans exhibit, albeit in a more limited context. The major technology underlying AI, machine learning (ML), is useful for engineering such autonomy, as it can learn from external data input, either with direct human oversight or without. In developing these highly useful technologies, knowledge from human factors and ergonomics (HF/E) can be of great use, especially to designers charged with the difficult task of dovetailing humans and machines in complex systems built to navigate sometimes chaotic environments. Technology serves as a greater extension of human ability each year, and optimal performance still results from hybrid human–machine teams (Figure 1).

Automation, autonomy, and AI are all distinguished by self-direction, and indeed it is arguable that these terms are synonymous in intent. As such, in the present chapter we will refer to them collectively under the moniker A^3 (pronounced "A cubed" /A kyübd). Herbert Simon (1965) named the technology of automation in his writing, which retains its ancient meaning "acting of itself" or "to rule one's self." The Latin roots of the word "autonomy" likewise relate to "making one's own laws," and so in this chapter we will use that term to indicate the degree to which a system or machine is under its own control. Automation, meanwhile, will be used to refer to the degree of replacing human work in a given domain or task (Figure 2). Using these terms, we can say that A^3 technologies, with varying levels of autonomy, automate tasks once exclusively performed by humans. These terms have been used quite interchangeably in the literature, patents, manuals, and other technical, scientific, and public discourse. Bradshaw, Hoffman, Johnson, and Woods (2013) explore these overlaps in terminology and surrounding misconceptions, by enumerating seven "deadly myths." These are: (1) The erroneous idea that autonomy is unidimensional, when in fact the term encompasses qualities such as self-directedness, self-sufficiency, and many more; (2) Numeric scales describing "levels of autonomy" are poor ways to scientifically ground these multiple concepts; (3) Autonomy is not a "widget," a specific technology, or a discrete property of the system; (4) A^3 systems as a rule are not truly autonomous, requiring human involvement on some timescale; (5) Full autonomy, when eventually achieved, does not obviate the need for human–machine collaboration; (6) Humans create systems incapable of collaborating with us at our own peril;

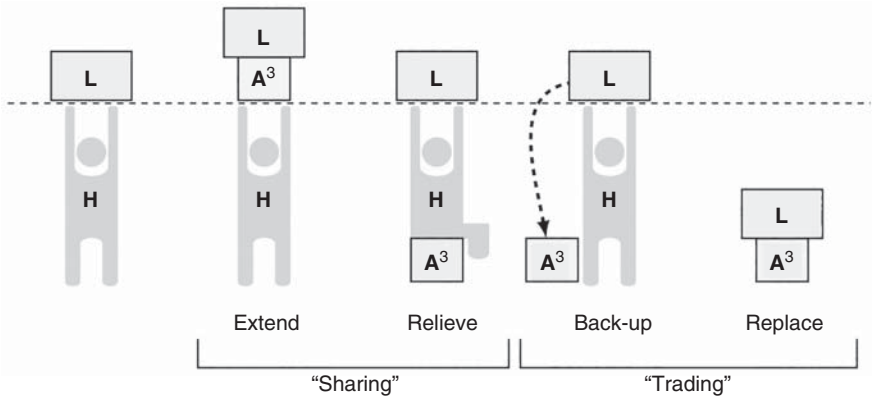

Figure 1 "L" represents the load or task, and "A^3" represent the automation, autonomy, or artificial intelligence (AI), in this metaphor for human–A^3 interactions in complex tasks. Relative to baseline human ability, at the left, A^3 can share the load, which extends human ability or provides humans periods of rest. A^3 can also trade the task with humans, perhaps serving as a backup. However, A^3 replacing humans will perform at a lower level than a human–A^3 team, due to naturalistic complexities. Best absolute performance comes from collaboration between humans and A^3 in most situations. (*Source:* Adapted from Sheridan and Verplank, 1978.)

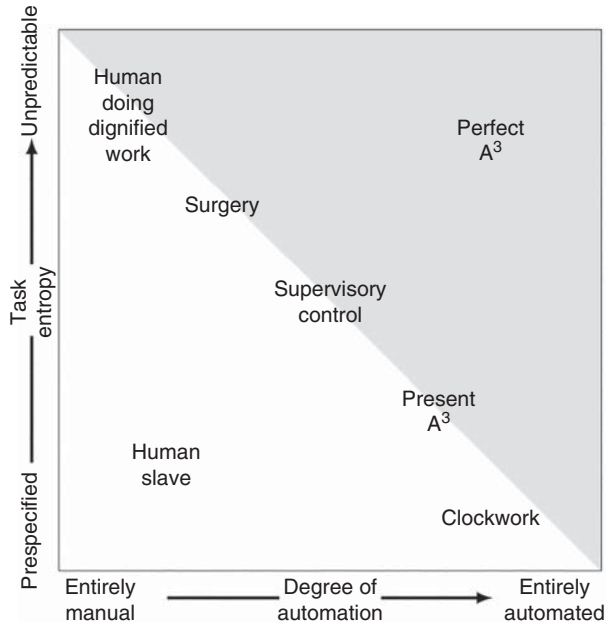

Figure 2 Classifying human and A^3 capacities across task entropy, which is often a function of environment complexity and degree of automation. While present systems are capable of handling prespecified tasks, and even moderate levels of task entropy, in more complex environments the best performance is achieved through supervisory control in which humans monitor and respond to incapacities in and failures of A^3, or in a collaborative control arrangement where a human- A^3 team manages the task together. Indeed, unpredictable environments where humans thrive may someday be a space amenable to A^3 action, and designers should begin now to consider where and if, in that eventuality, humans will find dignified work. (*Source:* Adapted from Sheridan and Verplank, 1978.)

and (7) Full autonomy is not likely possible nor is it universally desirable; and that autonomy is not necessarily humanlike and it does not replace human agency. These important points are necessary to understand human factors in A^3, and to design optimal, or even acceptable, human–A^3 systems.

The role of HF/E in A^3 design remains centered around the goal that A^3 self-action is ideal to provide maximum benefit to humans while increasing the likelihood of task success. The automation of human tasks employing A^3 also raises significant risks for individual users and society at large, related to the introduction of potential new sources of error, loss of privacy, data security risks, lack of explicability and transparency of algorithms, job replacement and loss, appropriate trust in and wise adoption of A^3, and surfaces issues of ethics and governance (Kearns & Roth, 2019). Human factors and ergonomics (HF/E) researchers and practitioners have for over 80 years faced the challenges and opportunities afforded by automation, defined as systems which perform functions previously managed by a human operator (Parasuraman, Sheridan, & Wickens, 2000). These years of work reveal impacts far beyond the replacement of humans, and shows rather that the introduction of A^3 has produced unexpected machine requirements and changed human roles (Woods, 1994). Indeed, the role of the human can move from exerting effort to supervising the effort of A^3, or *supervisory control*. As such, A^3 reshape rather than replace human effort. A^3 systems, with few exceptions, require human supervision and interaction, if only as the human experiences the final output. Sheridan and Verplank's (1978) classification of task factors and the suitability of tasks for humans or machines (see Figure 2) provides an introduction to human–automation collaboration, and also calls into question whether A^3 is always the answer (see Hancock, 2014).

A^3 can also help to overcome HF/E challenges, and lead to the successful employment of system automation (Coleman, 2019; Frischmann & Selinger, 2018; Shekhar, 2019). Furthermore, applications of A^3 in the context of human factors and ergonomics can be used: (1) to promote autonomy; (2) to predict human cognitive behavior; (3) to anticipate human physical states; (4) to analyze massive datasets of human measurements; and (5) to enable new human factors methods (Lau, Fridman, Borghetti, & Lee, 2018). For example, Canonico et al. (2019) used A^3 technology, team cognition, and collective intelligence to develop a new model for teamwork. Chollet et al. (2017) proposed a model to automatically extract sequences of non-verbal signal characteristics in order to develop a virtual recruiter that could express social attitudes. De Melo et al. (2012) developed an AI-based approach to create the ability to recognize emotions in a multi-agent system. These innovations highlight the importance of emotions in communication and human–machine

interactions. They give hope that A^3-assisted HF/E work can provide an understanding of complex human behavior-related design opportunities, and so address previously intractable problems.

The stakes for design of A^3 have never been higher, as illustrated by the 2018 death of Elaine Herzberg, the first recorded fatality due to an autonomous vehicle. Walking across a dark road, not at a marked crosswalk, Herzberg was struck by an Uber autonomous test vehicle which failed to identify the unexpected object as a pedestrian until 1.3 seconds before the crash (Griggs & Wakabayashi, 2018). The supervisory human driver was, prior to the incident, watching an episode of *The Voice* on a handheld device, looking downward, and not attending to the road environment. Both human and A^3 acted too late to avoid a collision. The full implications of this tragic event are still unfolding at the time of this writing. Testing of increasingly autonomous vehicles moves forward, and Uber's own report on the incident (Uber, 2019) has underlined the potential for autonomous vehicle technology to be safer than unassisted human drivers. Compelling design questions remain: how might the human be better integrated into the decision-action loop, and how might design equip future human–A^3 teams to prevent further tragic failures? There is, happily, good evidence that such design efforts can pay great dividends. For example, in a 2005 online chess tournament in which human players were allowed to use computer assistance, the best rated computer system, and the best rated human team, were both defeated. The victors, instead, were human players of intermediate skill paired with a less powerful computing system, but armed with a unique strategy for human–computing collaboration. "Weak human plus machine plus better process was superior to a strong computer alone and, more remarkably, superior to a strong human plus machine plus inferior process," said chess grandmaster Gary Kasparov of the match (Behymer & Flach, 2016).

The present chapter is written between "AI winters," (Floridi, 2020) times of decreased funding of AI technologies, indeed at a time of great optimism and investments in A^3 technologies. It is tempting, in a moment such as this, to imagine that carefully considered design is secondary to the ever-accelerating march of technological progress. Surely, A^3 can simply design itself? We argue throughout this chapter that, in fact, design is key to all A^3 technologies, and that the advent of autonomy which has no need for humans is not only unlikely, but likely undesirable. HF/E research and understanding serve as a vital foundation for building strong human–A^3 teams. Independent of other worrying indicators, such as the diminishing rate of increase in computing capacity that powers the deep learning techniques that underpin nearly all A^3 advances (Thompson, Greenewald, Lee, & Manso, 2020), we here submit that the design of effective human-computing teams for high-stakes tasks is likely to become of ever growing importance due to even greater capabilities of human-A^3 collaborative systems.

A^3 in collaboration with humans can provide great benefits, and has, as of this writing, had transformative impacts on domains including communication, commerce, design, global and extraplanetary transport, medicine, and security (Heer, 2019; Lau et al., 2018; Manyika, 2017; Mittelstadt, Russell, & Wachter, 2019; Raisch & Krakowski, 2020; Schaefer et al., 2015). Wang and Siau (2019) note that A^3 systems already interact with the real world and make autonomous or semi-autonomous decisions in both civilian and military domains including manufacturing and factory automation, automated vehicles, robots and drones, education, human resource management, cybersecurity, health care, finance, and management of hazardous environments. Vamplew et al. (2018) discussed the potential behavior of widespread A^3 systems, and both the benefits and challenges they bring, highlighting the issues of ethical, legal, and safety-based frameworks of human–machine interaction. We must keep in mind using A^3 to ensure the humane use of human work, rather than allowing machinery and automata to place humans in roles to which they are unsuited, use them as mere supervisors, or displace them altogether. As Norbert Wiener stated in his seminal 1950 book, *The Human Use of Human Beings*, "[machines should be] used for the benefit of man, for increasing his leisure and enriching his spiritual life, rather than merely for profits and the worship of the machine as a new brazen calf" (Wiener & Heims, 1989, p. 162). The proliferation of A^3 will likely be accompanied by changes in the way we view the roles of A^3 *and* humans, especially in emerging areas like crime and fraud prevention, social communication, brand management, customer services, software testing and development, human resource management, among many others (Coleman, 2019; Frischmann & Selinger, 2018; Kearns & Roth, 2019; Mittelstadt, Russell, & Wachter, 2019). Indeed, while some researchers argue that while A^3 threatens human control (Kearns & Roth, 2019; Mittelstadt et al., 2019), it is also redefining the identity of the humans who are exerting control (Coleman, 2019). We here postulate that A^3 will lead to re-engineering humanity as we know it (and see Frischmann & Selinger, 2018). The present chapter endeavors to provide tools from the HF/E literature with which to shape the development of A^3 with respect to our knowledge of human factors.

2 UNDERSTANDING HUMAN INTERACTION WITH A^3

HF/E research has long understood that, when designing the shift of control from human to machine or from machine to human, greater system autonomy actually requires greater consideration of the human operators' contribution (Bainbridge, 1983). It is indeed difficult to conceive of A^3 in which there is no human contribution. Such an imagined "completely autonomous" system would, at the very least, need to be set in motion by a human, or monitored for failure and replaced when no longer functional. Indeed, designers who imagine their systems as completely independent of humans often simply neglect to think on a large enough timescale. In extending human ability, A^3 systems require *supervisory control*, in which a human may (1) plan off-line; (2) teach the system; (3) monitor the system's actions; (4) assume control; or (5) learn from the actions of the system (Sheridan & Parasuraman, 2005). Such a system–operator response occurs in cycles: the human perceives the machine's state and takes an action, the machine senses the new state of the environment and itself takes an action. This cyclical interaction gives rise to the term *in-the-loop*, a characteristic of an operator who is actively exerting supervisory control. The problem of keeping operators in-the-loop is one of the principal design challenges of this moment in history, and unlikely to be solved in one "great leap" (Endsley & Kiris, 1995). Consider the complex challenges of keeping humans in-the-loop with other humans, also unlikely to be solved in one "great leap," and instead an unending labor of strategically building understanding. Humans and machines function as teams, and the need to keep such teams together and actively cooperating is a significant and exciting design challenge in its own right.

2.1 Human-A^3 Teams of Teams

Humans and A^3 working together are often referred to as teams (Groom & Nass, 2007; Jung et al., 2013). Crucially, these teams can be described as one human matched with one system (1-1), many humans facing a single system (n-1), a single human

facing many systems (1-n), or many humans facing many systems (n-n). Often, the ground truth depends on your frame of reference. For example, it is easy to conceptualize a single human driver "teaming" with an advanced driver-assistance system (ADAS) as a 1-1 situation, but stepping back to consider the larger driving public reveals that many individuals are interfacing with the same ADAS system (n-1). ADAS are often the aggregate of several individual systems, working in concert (Cades, Crump, Lester, & Young, 2017), and so an argument can now be made for this being an n-n situation. With the advent of constantly connected cars, the situation becomes again more complicated, as individual drivers may interact with a combination of systems present in their vehicle, and with remote systems. As the systems present in cars move beyond the safety-focused capabilities of ADAS, and toward assuming an increasing proportion of the driving task, the very concept of a human–machine team is challenged. On roadways with many automated or autonomous vehicles, systems more commonly communicate with one another (vehicle-to-vehicle, and vehicle-to-infrastructure) than communicate with the human, and that n-n system is therefore more tightly coupled with other machines than with the human in the vehicle. At what proportion of automation versus human control is the human no longer a significant member of the "driving team"? At what point does "teamwork" fall away, because humans and A3 are no longer defensively "working together" (and see Hancock, 2020a)?

Teams of humans can also interact with and benefit from integration with A^3, in n-n settings. Team cognition, in human–human teams is the binding mechanism which emerges from the interplay between members' individual cognition and process behaviors like coordination (Cuevas, Fiore, Caldwell, & Strater, 2007), and can in coordination with A^3 be augmented. For example, consider the design opportunities which exist around expertise coordination processes, such as asking, learning, sharing, and solving (Caldwell, Megan, & Jordan, 2019). In these, novices and experts might interact, exchanging or generating new knowledge, and A^3 would have the opportunity to enhance information exchange and updating, with positive impacts on team attention and shared situation awareness. Such hybrid teams of teams serve to amplify the already amplified intelligence of team cognition, providing decisive analytical and strategic benefits. Current generation implementations of this idea are rudimentary, at best, and there is an enormous possibility waiting to be tapped at the intersection of human and machine teams by enterprising A^3 designers. Indeed, the questions of what human intelligence is, and what it can be, are waiting to be reinvented.

2.2 Humans and A^3 Explore in Sequential Stages

A *perception-action cycle* consists of information being gathered through sensory organs, this being used to modify knowledge of the world, guiding decision-making, and culminating in directing or withholding action. This action changes the state of the world, and necessitates re-sampling, starting the *cycle* again (Neisser, 1976). While humans may make interactive changes, we are also able to react to changes stemming from outside agents or forces; changes within the environment are the impetus for re-engaging in the cycle (Gibson, 1969). Such cycles can be considered at many different timescales (K. Smith & Hancock, 1995), from the moment to the millennium, and potentially beyond. This cycle framework underlies useful psychological constructs, such as attention, memory, information acquisition (IA), situation awareness (SA), distraction, and multitasking. Multitasking, in this light, might be framed as concurrent performance of tasks requiring multiple concurrent perception-action cycles, each of which becomes more likely to experience delay and/or failure. Humans spend their entire lives engaged in such cycles, turning information about events in the outside world into insight as to what actions might elicit advantageous outcomes, and then directing selected actions toward impacting events in future cycles.

A^3 can, interestingly, also be understood through a cycle-based metaphor. A machine system samples through sensors, or likewise engages in the acquisition of information, and from this arises a set of processes which can be conceptualized as paralleling the human perception–action cycle. This is followed by analysis of information, with an option to display that information to a human teammate. Next, a decision regarding actions to be taken are rendered. Finally, the A^3 engages in implementation of an action. This, in turn, starts the *sequential stages of an automated system* over again (Parasuraman et al., 2000; Sheridan & Parasuraman, 2005, p. 93). In human–A^3 teams, supervisory control can occur at every sequential stage of an automated system, and each has the opportunity to be assigned to a team member: a computer or human (Parasuraman et al., 2000). To be clear, this is not a direct parallel in all engineered systems, but it is a useful and time-tested approximation in considering system design (Broadbent, 1958; Rasmussen, 1986).

In attempting to integrate the perception–action cycles of humans with the sequential stages of A^3 systems, a designer must consider human tasks that fit with the machine process. Humans already follow an ingrained set of linguistic turn-taking behaviors, or *antiphony*. The antiphony framework leverages this inherent call-and-response of language, and allows engineering-oriented, time-based expression of interplay between human and machine at the task, and subtask, levels (Sawyer, Mehler, & Reimer, 2017). In a 1-1 interaction, for example, a designer might determine when the human operator was engaged in perception, cognitive processing, and decisions, and when they directed action toward the system, the period when the system was processing, and when it was producing multimodal cues for the operator. Processing by the human operator, or the system, often occurs in parallel with the other party's action, and so the period of operator latency, and system latency, when either party pauses for processing, can be as low as zero. Likewise, the duration of system and operator action can be as low as zero, as inaction can occur, as well as action interruption by the other party. The cycle (Figure 3) can start from any step, and provides a framework for predictions of time, and to understand the results of measured time. For example, in evaluating and improving the effectiveness of an A^3 email security system, the finding of long delays in operator latency might prompt different questions and attempted design interventions than the same delays in system action. A savvy designer will think about opportunities to dovetail the perception action cycles of human operators with the sequential stages of automated systems, toward functional, integrated teams.

2.3 Human-A^3 Team Function Allocation and Levels of Automation

As in all teamwork, a constant challenge in human–A^3 teams is delegation. In cooperation, there are advantages to allocating tasks to the team member most capable at the current moment. As a result, the contributions of a human operator are helpful to consider in relation to present and projected near-term A^3 capability. For some 70 years (Fitts, 1951), tables detailing the superiority of humans and machines by task type have been used for *function allocation*, the assignment of tasks in a human-machine team. A related approach is to consider the differing proportions of human and machine involvement in the overall task ecosystem, and construct *levels of automation* (LOA), which designate which party does what across multiple

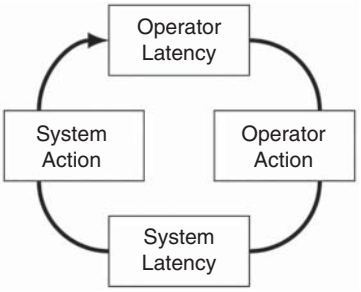

Figure 3 The Antiphony Cycle, which uses the call and response of language as a framework to integrate perception action cycles of humans with the sequential stages of A^3 systems, can be used to plan a design, or analyze data collected from human-A^3 interactions. (*Source:* Adapted from Sawyer et al., 2017.)

levels of autonomy. Many tables of function allocation and levels of automation have been proposed over the years. The analytical approach to human–computer cooperation, one of the oldest taxonomies, was proposed by Sheridan and Verplank (1978) and revised by Sheridan (2002). A more concise taxonomy with four automation levels was introduced by Endsley (1987, 2016). Riley (1989) discussed a novel taxonomy as a 2-D matrix with rows corresponding to the level of automation and the columns to intelligence levels. Milgram et al. (1995) proposed a taxonomy with five LOA, considering the different roles a human operator could play in remote manipulator systems. Draper (1995) developed a taxonomy that combines human operators with machine control in teleoperation, with five automation functions carried out by the human operators and the other four allocated to the machine. Kaber et al. (1999) identified numerous LOA combinations which were not included in the former taxonomies, with a detailed description of "who" (human or system) is supposed to do "what" (task) at each level as compared to the previous hierarchies of degrees of autonomy. Lorenz et al. (2001) present a more compact taxonomy than the previous ones, consisting of only three LOA in the context of automation support. Function allocation taxonomies are also built specific to domain. Clough (2002) described a four-level automation taxonomy with a specific application in unmanned aerial vehicles. Proud et al. (2003) presented a taxonomy of automation with eight different levels of autonomy scale to fit the tasks encompassed by a function type, i.e., observe, orient, decide, or act. Finally, Fereidunian et al. (2007) presented an extension of the LOA taxonomy introduced by Sheridan (1992b), adding an eleventh automation level. Indeed, many more such taxonomies exist, and a designer in a specific domain might be unsurprised to find that another group had already considered the priorities of delegation in that context.

In specific domains, such as automated driving, such levels can be quite specific and the basis for both technological design and regulatory policy (SAE International, 2016). It is extremely important to recognize the inherent complexities of designing such levels, and that applications relative to their definitions and implementation may be revealed along the path of their design and development. For example, the Society of Automotive Engineers' (SAE) five levels of driving automation (LoDA) defines "conditional driving automation," as a self-driving vehicle performing "the complete dynamic driving task (DDT) within a limited operational domain," for example 'highway driving'. The word *complete* notwithstanding, the driver is expected to respond to any and all automation-issued requests to take over control in a timely manner, which may be difficult or impossible as considered in light of human factors research. This ambiguity does not mean that the SAE LoDA are "wrong," but does reveal an opportunity for clarification of the manner in which A^3 and a human share and trade the task of driving (as in Inagaki and Sheridan, 2019).

Mission-critical tasks with no clear pattern of superiority can be better understood through the straightforward analytical approach provided by Sheridan and Parasuraman (2005). Indeed, this probability-based signal detection approach makes it ideal not only for one-time understanding of deciding whether an A^3 or human operator is a better fit for a task, but for longitudinal functional allocation analysis of tasks through changes in context, requirements, and technological progress as well. When the answer to function allocation is "it depends," the answer is often to move beyond static allocation of functions to dynamic allocation of function using adaptable or adaptive automation (Sarter, Woods, & Billings, 1997; Scerbo, 1996; Sheridan & Hennessy, 1984). Dynamic function allocation has the advantage of providing automation which is robust to the changing demands imposed by shifting environments and human capabilities. For the human operator, this means improved performance, including control during anticipated (Harris et al., 1995) and unanticipated (Hoc, 2000) events. The idea of longitudinal function allocation can be nested with the concept of dynamic allocation of function, recognizing that identifying which agent is assigned what task is, in human–A^3 interaction as well as in human–human interaction, best viewed as an ongoing process. Of course, optimal delegation cannot always be mathematically derived, and may indeed be driven by convention, custom, and individual preferences of humans, which are likewise important. Human desire for control should be factored into decisions regarding the application of A^3. Humans require engagement in the tasks we are to undertake, and even to have some fun, with due regard to safety and well-being (Hancock, Pepe, & Murphy, 2005).

While such heuristic tools seem simple, only asking a designer to look through the functional requirements of the task, and capabilities of humans or machines, and then to assign them to a human or A^3 may not be enough. All of the possible combinations of human and A^3 collaboration across even a single task form a big space, and it is difficult to project exactly how the human–machine system will behave in a naturalistic environment. There is a risk of placing humans in "gaps" where automation is incapable of fulfilling the needed role, or placing A^3 in places where human capacities are limited. Consideration must be given to the shifting interdependencies that underlie complex activities with multiple agents engaged in such tasks. Engaging in function allocation without considering interdependencies can result in the human performing only a fractured collection of subtasks too difficult to automate (Bainbridge, 1983), or worse, different fractured sets in different situations. Therefore, especially in high-stakes activities, testing in the environment the system will be used in, or a reasonable

facsimile, is vital. In determining task assignment to a human or computer, Fitts' List variants (Fitts et al., 1951) and LOA are best used to form a starting point for consideration, with deeper testing-grounded analysis playing an absolutely vital subsequent role. No heuristic rule set can perfectly serve the larger goal of identifying opportunities for a hybrid human–machine team to place team members in ideal roles with respect to achieving the target goals (Hollnagel & Bye, 2000; Sheridan & Parasuraman, 2005).

Finally, taxonomies and LOA fundamentally ignore the great variability in human beings. While designing for "the general public" is often necessary, A^3 provides a unique opportunity to individuate, to allow the system to accommodate the individual. Indeed, if there is any great opportunity in the advent of increasingly flexible machines exercising greater autonomy, it is that they might begin recognizing and tailoring their interaction to each of us, much as we do to one another. In order to understand these opportunities, designers are advised to build an understanding of how humans differ from one another. One useful way to conceptualize these differences is through traits, states, and the focused application of learning, training.

2.4 Human Traits, States, and Training

Broad differences in human ability, both between individuals and within the same individual over time, reveal another level of design: the question of whether A^3 or a human should be on point for a task depends both upon the person, and on the moment. The capabilities of a racing driver and the average driver differ markedly, and a drunk driver's capacities fall even further down on the curve of potential performance. More complex still is the possibility for a person who is an excellent driver in one moment to become a danger to themselves and others in the next. One's *state*, such as being well rested, fatigued, drunk, focused on driving, or cognitively preoccupied, is a rapidly changeable factor in skill and capacity. *Traits*, meanwhile, are relatively stable factors which change only over prolonged periods of time. Age, height, spatial reasoning capability, and disability are all trait factors which persist across time in an individual. *Training* spans the intermediate temporal space: individuals can improve their capacities through dedicated practice or learning, sometimes in the space of hours. States, traits, and training interact to determine capabilities in the moment, and both the design of A^3 and the configuration of adaptive systems need to take these into account on the appropriate timescale.

State factors are rapidly changeable parameters, often highly related to actions in the moment. A^3 systems can respond best to variation in human states through a combination of monitoring the human, referred to as *state detection*, and making available their own state, through system transparency (J. Y. Chen et al., 2014). For example, sleepiness can significantly interfere with cognition, impair decision making (Hockey & Wiethoff, 1993), increase reaction time, and negatively impact vigilance (Philip et al., 2005). A driver's state can be detected by A^3 through driver-facing camera approaches monitoring eye and eyelid information. This state estimation can be used to provide alarms or interlock functionality (Sommer & Golz, 2010) and change the level of automated agency accordingly. The separate state of vigilance decline, which can reduce ability to detect signals in as little as 30 minutes (J. F. Mackworth, 1969, 1970; N. H. Mackworth, 1948), and which is common in supervisory control of A^3 (Sheridan, 1975, 2006, p. 200) has no equivalent camera-based detection mode, and must be detected by EEG or other less convenient biosignals (Greenlee, DeLucia, & Newton, 2018).

Any state detection technology should be evaluated in terms of its ability to identify relevant human states, in naturalistic environments, and under normal operational conditions. The ability to identify human states in the lab, while a useful initial data point, is often not indicative of such real-world performance. It is also important to be certain the state can actually be deduced relative to the context, and that the data upon which assumptions were built is sound. For example, direction of the eyes can be detected through computer vision approaches, but its use as state detection is controversial because of the difficulty of matching eye direction to environmental information (Wolfe, Sawyer, & Rosenholtz, 2020). What is visually present in the environment matters, as changes there can rapidly increase the demands on a technology supervisor. Likewise, irrespective of eye position, cognitive focus can be diverted away from the focal task by distractions (Lavie, 2010), "trapped" through tunneling into a task that occupies cognitive resources (Simons & Chabris, 1999), or spread thin among multiple concurrent tasks (Salvucci & Taatgen, 2008, 2010), some of which may not even relate to the task at hand ("Did I turn off the stove?"). The malleable attentional resources theory (Young & Stanton, 2002a, 2002b, 2006) suggests that attentional resources can be reduced by periods of low load, and that when there are rapid increases in attention demand from the environment, which can leave a supervisor with insufficient cognitive resources to detect a change, or respond to a pressing situation, especially if there is a silent failure of automation that is not detected.

Most, if not all, new approaches to state detection in driving can be revealed to be either opaque machine learning models, or derivations of the seminal work of John Senders (see Eisma, Hancock, & de Winter, 2020), and indeed the former may yet be determined to be the latter. Foundational findings hold stubbornly true in today's attempts to integrate state detection into A^3. Direction alone is not sufficient to understand what is being looked at and how it affects the human, and so no A^3 system, or subsystem, can accurately predict human attention or cognitive load without having access to both gaze direction and the contents of the environment. Finally, consider that state detection is itself an A^3 system, and so the provenance of the underlying data and assumptions is important. For example, recent evidence suggests that much of the eye tracking data used to build models of driver distraction and understanding of in-vehicle behavior may in fact be systematically flawed (Jansen, van der Kint, & Hermens, 2020), a serious consideration for designers using such research to underpin state detection or design decisions in production settings. Visual behavior-based state detection is, of course, but one exemplar type, and many other types of state detection fail similar logical tests in certain contexts. State detection, as a capability of A^3, is integral to interaction with humans, and increasingly broadly applicable and effective, but it must be used with great caution.

Trait factors can be described as unchanging or slowly changing attributes, such as age, personality features, or physical or cognitive capacities or disabilities. These factors can define how A^3 should be designed to best work with stable categories of human users, compensating for weaknesses, or leveraging strengths. Cognitive and physical capacities decline with advancing age (Kline et al., 1992; Waller, 1991), and inexperience or the greater risk-seeking orientation of young people (Jonah, 1986; McKnight & McKnight, 2003) make age an important consideration in the adoption of A^3 technologies, or at least a proxy for mental and physical capabilities and risk-seeking or risk-averse orientation. Children, adolescents, and teenagers also have different physical and cognitive capacities compared to adults, and thus the design of technology should take into account the intended user base. In actuality, A^3 systems may have more to offer older users than to people in their physical prime, as they can provide capabilities an individual can no longer provide for themselves, and therefore an extension of the time during which an individual can age in place (J. Miller et al., 2018).

Training, and the underlying human capability of learning, provide a rapid and somewhat more permanent way to affect performance. Military personnel, industrial operators, pilots, and professional drivers have extensive training on how the automated systems they use operate, their limits, and their capabilities. Training helps form an accurate and accessible mental model (Halasz & Moran, 1983; Johnson-Laird, 1980; Norman, 1983; Richardson, Andersen, Maxwell, & Stewart, 1994; Wilson & Rutherford, 1989) of the system, perhaps one more comprehensive than one formed through exploration alone. Kieras and Bovair (1984) experimentally explored the development and use of mental models using an entertaining *Star Trek*-themed experiment, finding that people with training that helps form a model of the device's behavior showed faster learning and that the model can help users to infer operational procedures. Other research showed that providing users a conceptual model of a programmable calculator's memory system (Halasz & Moran, 1983) aided in novel problem solving, while not improving performance in routine problem solving. Inaccurate mental models may cause improper use of systems, both simple ones like a home thermostat (Kempton, 1986, 1987), and more complex systems such as computed tomography (CT scanners). Barley (1988) notes that users of complex systems can create "anthropomorphic" representations of systems, ascribing them humanlike agency where it is only the interaction between system and user behavior that generates what could be considered fanciful mental models. Most A^3 systems, at least as of 2020, don't have "bad days" or actively plot to thwart your aims. On a more functional note, A3 systems certainly have bad contexts, in which their ability to contribute to a task may degrade to the point where they thwart the aims of the user.

If users are not properly trained in the use of a complex system, they may develop strategies for clearing errors or resolving problems that do not correspond to reality, potentially reducing efficiency or compromising safety, or that don't actually solve the problem at hand. It is worth considering that this mental models research was conducted in the 1980s, with relatively simple systems compared to the much more complex A^3 systems now common in our workplaces, homes, vehicles, and pockets. A^3 and training in its use have evolved, and not in a way to be simpler for the designer.

Designing training and instructional materials for usability are serious concerns for A^3 development and deployment. Outside of the rigid requirements of professional, legally required, and conscription-based training, it can be very difficult to entice users to train at all. One reason for this is the impermanence of systems that are constantly under revision, and for all the benefits of software-updatable systems, they do not update the human and so can render training instantly obsolete. This reality has given rise to the prominence and importance of technologies which "just work" or are "intuitive." In reality, many such technologies are built on sets of conventions we have learned over our entire lives, a form of iterative training. For example, the iPhone was introduced over a decade before this writing. It has since been the most common smartphone by several metrics, so any claim of inherent "iPhone usability" must be measured against the fact that a generation has literally grown up with the design and interaction language of this product. The barrier to entry that many A^3 systems face in finding "intuitive" behavior that users will engage with can also be framed as the challenge of finding behavior close enough to existing conventions to allow for adoption without undue investment in deliberate training.

Designing systems to be explored by users should not be relied upon as a replacement for training or documentation, exploration should ideally be combined with training. To that end, A^3 systems should include elements that help guide user behavior, such as task-guidance wizards or intelligible documentation written for the expected level of user knowledge, and facilities should be included to avoid irrevocable or hazardous action. Where it is feasible and reasonable to explore an interface's affordances, users will do so and probably should; and design should support them in that endeavor.

While these categories of traits, states, and training are useful, some areas of interest are obviously a blend of all three. Personality factors and experience with technological systems both have a role to play in how people use technology (Davis, 1989) and trust in technology (Hoff & Bashir, 2015). A meta-analysis by Hancock et al. (2011) found that the largest determinants of willingness to trust a robot is its prior behavior, with personality factors playing only a minor role. Propensity to trust as a trait (Frazier, Johnson, & Fainshmidt, 2013) may have a significant role to play in a human–agent relationship, influencing use or adoption, especially in situations where there is minimal prior experience with the system to draw on. While the intersection of personality with user behavior is undoubtedly one of the most challenging aspects of modern A^3 development, machine learning approaches have shown ability to parse this complex intersection (De Melo et al., 2012). We see great opportunities for designers willing and able to tackle this complex opportunity.

2.5 A^3 Complementarity in Systems Architecture

From a systems architecture point of view, integrating humans with A^3 is in part the art of being certain of the relative strengths and weaknesses of each party to complement one another. Jarrahi (2018) discussed human–automation interactions with attention to the complementarity of humans and A^3, in the context of organizational decision-making processes, and has suggested that while machine learning models align with the analytical decision-making approach, they are less viable in unpredictable and uncertain situations—humans having at this time greater cognitive flexibility. This fragility when faced with the unexpected is a fundamental weakness of A^3, and is correspondingly a strength of humans. Therefore, it follows that machines take care of mundane and predictable tasks, while humans focus on creative and uncertain tasks, but this is not always advisable (Takayama, Ju, & Nass, 2008). By considering capacities in relation to uncertainty, complexity, and equivocality, appropriate roles for humans and machines can be found within organizational decision making (see Figure 4), including collaborative arrangements using humans and A^3 as teammates.

Under this arrangement, A^3 assists human decision-makers in terms of uncertainty with predictive analytics, and by detecting relationships among hidden factors in a system. Intuitive approaches used by humans under conditions of high ambiguity are optimally combined with the superior speed of A^3 in collecting and analyzing information, resulting in better handling complex problems. Furthermore, A^3 can facilitate human decision-makers in handling equivocal situations and resolving relevant conflicting needs. However, resolving equivocality is at this time definitively the responsibility of human actors. Jarrahi (2018) concluded that humans offer a unique and irreplaceable intuitive approach for handling uncertainty and equivocality, while A^3 provides support to extend humans' cognitive abilities. Agrawal et al. (2019) argue that the proper design of A^3 interaction requires distinguishing the concepts of prediction and judgment, where prediction reflects the information about the expected state of the world, while judgment relies on factors that are difficult to describe and therefore codify (see Figure 5). Role and types of judgment might indeed be a function of whether or not prediction improves the value of judgment. For example, without adequate prediction, judgment can direct decisions toward riskier actions, while with enhanced prediction capabilities, judgment can direct decisions toward safer actions.

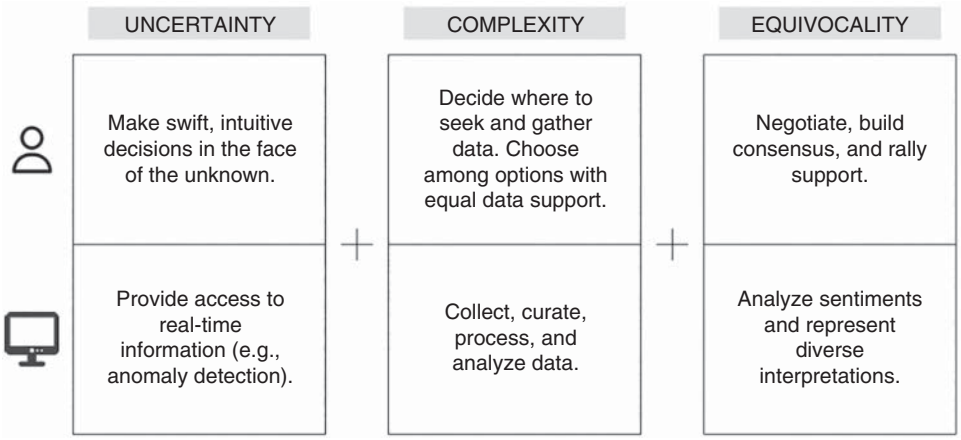

Figure 4 Humans and A^3 have different strengths and weaknesses with regards to operating under uncertainty, dealing with complexity, and resolving conflicts (termed equivocality by Jarrahi). Humans and A^3 can work in a complementary capacity in decision making situations, collaborative systems offering the greatest total envelope of performance. (*Source:* Adapted from Jarrahi, 2018.)

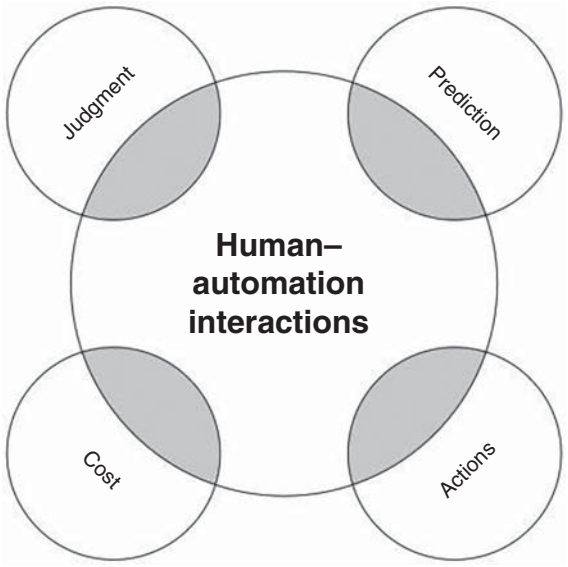

Figure 5 Augmenting human capacities with machine capabilities in prediction, and cost control, and leveraging human abilities in judgment and action selection can yield optimal outcomes in collaborative decision-making and control situations. (*Source:* Adapted from Agrawal et al., 2019.)

Architecting human-A^3 systems therefore requires balancing not only aspects of the human and machine, but thinking through scenarios in which these aspects hold weight and affect outcomes. There is no "free lunch," and every interaction of A^3 with a human operator has a cost, in terms of capacity of the system, but also in terms of human attention, workload, and potentially task success. This is true, even when human–A^3 interaction is beneficial, even when the intention is not to distract, and even when users like the interaction (see Sawyer, Finomore, Calvo, & Hancock, 2014), and so a designer must carefully consider human performance in A^3 design. What we have here focused on is a number of underpinnings of human interaction with A^3, many more surely exist. Indeed, insight as a designer into hybrid human–A^3 systems rests on the foundation of the design implications of psychological, biological, software, and hardware factors. A^3 systems are most effective when human-centered, designed to leverage a clear understanding of HF/E, to address human capabilities relative to realistic levels of training.

3 DESIGN OF HUMAN-CENTERED A^3 SYSTEMS

The design of A^3 can be described as the engineering of affordances for shifts in control over the course of tasks. A^3 and humans share responsibility for the task at hand, and in pursuit of that goal regularly take charge, or hand off control authority. A continuum can be imagined, from full manual control by a human user, through to full autonomous control by A^3, with multiple levels of shared control between those poles (see Figure 6). The arbiters of success in the shifts between these levels of shared control include the capability of the operator and the system to initiate and respond to changes in their partner's level of agency, to the quality of information transfer, and also the resilience of both actors to failures in transitions on the part of the other. This resilience to unexpected outcomes is one of the greatest current challenges facing designers of A^3. Greater degrees of change in level of autonomy, for example, from fully autonomous to complete manual control, raise the risks of failed handoffs, and changes between levels within the shared control regime can result in operator confusion, or losing track of which agent has authority over various control functions.

Whenever a human and A^3 interact, there must be one or more modes of interaction, or channels through which information flows. Multimodal communication with A^3 has more opportunities to be a bidirectional process than might be possible with less agentic systems. User exchange with a tablet computing device, for example, might feature information flowing from tablet to human via (1) vision in the form of a high-density display and perhaps additional light indicators, (2) audition in the form of audio output, and (3) touch in the form of haptic vibration. Simultaneously, information flows from human to tablet, and associated connected systems, via (1) manual touch, (2) manual movement of the device, (3) auditory voice commands, and (4) possibly machine vision. Humans receive information from A^3 via sensory channels, and then perform cognitive interpretation, render a decision, then plan and execute actions that direct information back. More traditional

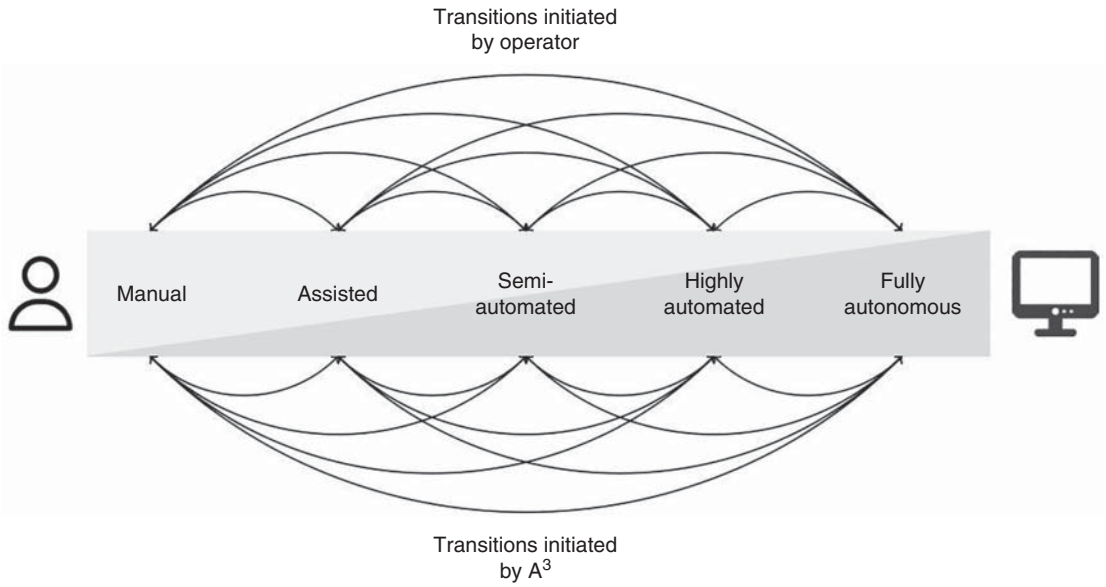

Figure 6 Level of autonomy can be considered on a spectrum from fully manual control to fully autonomous control, with multiple levels of shared or switched control between them. Transitions in level of automation can occur between any two levels, initiated by both human operators or A^3 systems. The greater the degree of change in level of autonomy, the higher the risk of operator confusion, automation surprise, or failed handoff. (*Source:* Adapted from Flemisch et al., 2008.)

systems have historically handled input deterministically: press a button to elicit a predetermined response. With A^3 systems, received information may be handled in a nondeterministic way. One can conceptualize such successive exchanges of information as a dialogue, and indeed the call and response of antiphony is a useful way to imagine such interplay. On both sides of this exchange, quality of information is a strong arbiter of a successful interaction.

In quantifying the success of moving information, the concept of situation awareness (SA) is an excellent way for designers to conceptualize user perception of A^3 interaction, comprehension of associated information, and projection of future actions (Endsley, 1995). Loss of situation awareness can lead to an "out-of-the-loop" situation, where an operator loses track of what an automated system is doing—and not doing (Endsley & Kiris, 1995). A related threat is mode error, where an operator does not know what an A^3 system is configured to do (Sarter & Woods, 1995; Woods, 1994). In system design, interfaces should make it difficult to inadvertently change modes, and provide clear feedback regarding mode changes. Judgment, prediction, actions, and associated costs must be constantly monitored by both humans and A^3, and SA on the part of the human, and system representation of human SA by A^3, can be vital. This is important when considering A^3–human interaction in dynamic, complex environments, and even more so when outcomes can have significant costs including injury or death. When the question is division of attention between an A^3 system and the environment, designers can enhance multitasking performance while keeping distraction minimal by ensuring that situation awareness is maintained by human operators (Skrypchuk et al., 2019). Pre-SA consideration of acquiring the information needed in order to respond may also be helpful, and the ideas of Information Acquisition (IA) specifically address how visual information is acquired (Wolfe, Sawyer, & Rosenholtz, 2020). This can help designers to understand how acquisition of information scales, and the process by which users construct and update their representation of environments. Design based around such understanding can strongly impact how information moves from A^3 to humans, ameliorating failure modes like distraction or information overload.

3.1 Information Design

In pursuit of successful responses, how can information passed from A^3 to a human be designed optimally? Here we will discuss two applied strategies for enhancing multimodal interaction: (1) maximizing interpretability; and (2) maximizing synchronicity. Interpretability differs by mode of information, but at present the most common modality for interacting with A^3 systems is visual. Displays of all sizes (Hancock, Sawyer, & Stafford, 2015) are our portal to the majority of A^3 systems, and through them, as of this writing, Americans spend around 5 hours a day on email alone, albeit not continuously (Chung, 2019). Here, they interact with increasingly sophisticated A^3 intended to monitor and co-manage the inbox. They also spend 2.5 hours on social media (A. Smith & Anderson, 2018), where A^3 matches them with contact-produced posts from friends and family, and sponsor-generated content. A^3 underpins searches, and moderates chat, finishes outgoing messages, and manages notifications for those inbound.

In the visual modality, the written word is still the most common form of interface, and of content consumed (Chung, 2019; A. Smith & Anderson, 2018). In messages from A^3 intended for rapid comprehension, attention to fundamentals such as typography and formatting provide substantial performance increases, and in contexts like driving, measurable safety benefits (Sawyer, Dobres, Chahine, & Reimer, 2017, 2020; Sawyer, Wolfe, et al., 2020). In longer-form reading, individual differences may be much greater, but potential gains are also significant: optimal information design saw boosts of over 100 words a minute for readers of short passages of text, enough to read an additional 10 pages in an hour with equivalent comprehension (Wallace et al., 2020a, 2020b). The principles of legibility in textual information certainly have parallels in visual design leveraging iconography. Indeed, text is not alone in its ability to convey complex information. For example, Chernoff faces (Bruckner, 1978) leverage fast

human processing of facial features to deliver complex information rapidly, and other compelling iconographic schemes for representing multivariate information have also been put forward. In discussing rapid information communication from A^3, therefore, it is important to consider how well biological systems achieve these goals. The enormously effective displays of the human face and body language are aspirational examples.

Synchronicity refers to matching A^3 output to human temporal rhythms of communication, such as antiphony. Humans have traditionally borne responsibility for this when interacting with machine systems, and indeed training in interacting with A^3 often explicitly focuses upon when to communicate. However, A^3 provides new opportunities to engineer not only system response, but even human perception of time. Humans are constrained by physiological and cognitive limits: a simple reaction takes approximately two-tenths of a second (Johansson & Rumar, 1971; Lerner, 1993), and more complex decision processes increase that time. Additionally, aging, fatigue, or distraction can markedly influence cognitive processing and reaction times. Fortunately, automation can stand in for human agency, both cognitive and motor. Systems such as automated emergency braking (Coelingh, Eidehall, & Bengtsson, 2010) can react faster than humans, and can reliably exert maximum force to more quickly slow the vehicle. A^3 systems can synthesize some forms of information more quickly to augment or even replace human capacities, although fully autonomous systems must be employed with care. Without oversight, errors may occur at a rate the human cannot check. Time pressure can sway human decision making, influencing people to filter information (Maule, Hockey, & Bdzola, 2000), focus on negative information, and even make riskier decisions (Ben Zur & Breznitz, 1981). A^3 can be employed to help humans to filter or focus on more relevant information, select from a narrower range of choices that have been preselected, or provide independent analysis of information when under time pressure, even if humans maintain ultimate control authority.

Finally, in considering synchronicity, it is important to realize that human perception of time is itself quite malleable. Hancock (2018) describes human perception of time as a construct, a product of human design. As a result, time perception can be manipulated through changing the physical and cognitive environment. Designing the information provided by an automated system can shape human perception of time passing, as anyone who has watched the status indicators of a system move toward indicating completion, or waited for a kettle to boil, will appreciate. While it may be less than forthright to design communication to manipulate perceptions of a system's capabilities, it may be warranted in the service of reducing user frustration, encouraging appropriate use of the system, and providing ability to build a usable mental model, ignoring this problem from a design perspective could lead to futures in which humans and machines struggle with fundamentally incompatible conceptions and perceptions of time (Hancock, 2020b).

3.2 Balancing Multimodal Workload

Humans are not infinite in their capacity, and so it is important to design A^3 to consider the workload required of the human, and to attempt to avoid overload and subsequent failure of the human–A^3 endeavor. Both humans and A^3 taxed beyond ability do not necessarily fail immediately, but may experience a pattern of successive worsening failures termed *dynamic instability* (Hancock & Warm, 1989). While failure is most impactful in situations with a high cost, it is worth considering that even low-stakes moments of overload and subsequent failure contribute to disengagement from, abandonment of, and eventual failure of A^3 systems. Operators, users, and customers will not engage with technology that demands too much.

Operators or users are often placed in a supervisory role where they will need to respond to changes in environmental conditions or system faults, sometimes with and sometimes without alerts. "Silent failures" (Louw et al., 2019) present great danger, as attentional resources may be focused elsewhere, and without an alert, the failure may go unchecked or unnoticed by a human. A famous example of this is the "invisible gorilla" that is routinely overlooked in video recordings of basketball games (Simons & Chabris, 1999), websites (Gelderblom & Menge, 2018), and x-rays (Drew, Võ, & Wolfe, 2013) even though visually fixated upon. In situations of high cognitive load, it is easy to miss even something as surprising, or out of place, as a gorilla in the center of the visual field. It should therefore come as no surprise when people, focused upon other details, routinely miss changes in system state or in the environment. While ultra-reliable systems and self-checking components can reduce the threat from silent failure, they cannot be eliminated. A wiser solution may be encouraging appropriate reliance and encouraging greater attention and vigilance, of course considering human limits in maintaining vigilance and situation awareness.

The most important design note for A^3 architects concerned about workload may be that the same amount of demand across multiple multimodal channels can produce lower overall subjective workload, as compared to moving the same demand over a single modality (Wickens, 2002). A^3 further provides opportunities to reduce workload in operators by filtering information provided. Rather than monitoring many indicators, an operator can attend to a supervisory system, or interact with an interface that controls many subsystems. The risks of this approach are that a supervisory system itself can fail, oversight in design could compromise its performance, or attention can be directed away from this system's interfaces (Sheridan, 2006; Sheridan & Verplank, 1978). Cognitive focus can be controlled voluntarily, or by an external agent (Wickens, Santamaria, & Sebok, 2013; Wickens et al., 2016), but if A^3 is tasked with controlling or directing attention, it must be highly reliable, be appropriate for the environmental conditions, and use appropriate modalities for signaling. In this regard, use of modalities by A^3 can be strategic, assuming A^3 has adequate awareness of human state. For example, A^3 with awareness that the operator visual channel is occupied, might issue a haptic or audible alert, or issue a visual alert in the field of the user's peripheral vision. This has strong parallels in human–human interaction, where an individual looking away might be engaged with a tap on the shoulder, or hearing their name spoken. The error rate of systems is relevant here, as systems prone to false alarms may encourage users to become attuned to ignoring alerts, and therefore miss true alerts or delay responding to them (Wickens, 2013). Systems more prone to misses may encourage greater human vigilance (Fu et al., 2019), but automation misses may have drastic consequences if unchecked.

The *prevalence paradox* (Sawyer & Hancock, 2018) is one common form of A^3-induced operator propensity to miss signals. Humans need a certain amount of information to understand the world, and so rare signals are more difficult to find, even when taking into account their low occurrence. Indeed, these failures-to-find represent diminished human capability to both detect and respond, a recognized, and often deadly problem in contexts including air traffic control, baggage screening, and cancer diagnosis. A^3 often work to reduce the work of the human by handling events that occur, but in doing so they are removing information the operator needs to understand the world, and so making humans in their team increasingly likely to fail in detecting and reporting remaining attacks. When A^3 successes become the seeds of human failure, what implications for human–machine teaming must arise from this prevalence paradox?

In vision, the "attentional spotlight" theory (Norman, 1968; Posner, Snyder, & Davidson, 1980) describes attention as a single cognitive focus that can be steered toward or attracted to stimuli, and recent understanding shows that it is joined by highly sensitive peripheral vision encompassing the remaining visual field. Other research has introduced a concurrent performance model (Navon & Gopher, 1978; Wickens, 2008) to complement the sequential processing models. The Noticing-Salience Effort Expectancy Value (N-SEEV) model (Wickens, 2013) and the threaded cognition models (Salvucci, 2013; Salvucci & Taatgen, 2008, 2010) address both the sequential and concurrent processing paradigms. Wickens et al.'s (2016, 2017) research on attention switching has led to the development of the Strategic Task Overload Management (STOM) model, which predicts that in a condition where there are insufficient attentional resources to meet demand, a load-shedding process will be initiated, where preferential focus is placed on high priority, lower difficulty, engaging, and highly salient tasks.

Task switching is not only a problem for pilots and system operators: the amount of time computer users spend on a focal task before switching between windows or tabs is on average only a few seconds (Yeykelis, Cummings, & Reeves, 2014, 2017). Research by Reeves et al. (2019) investigated task switching in computer use, finding that the ability to switch the content displayed on a single screen makes possible the "fragmentation of experience" in a way not possible with other media such as books, or even other types of displays. Yeykelis et al. also found evidence for rising autonomic arousal anticipating switches from "work" to "entertainment," but not a corresponding rise when switching away from entertaining to less entertaining stimuli. This reinforces the claims of engagement as an important component of the STOM model where it is more difficult to cognitively release from engaging tasks, even in critical situations (Horrey & Wickens, 2006).

Task focus has been commonly measured in laboratory settings using eye tracking, often in concert with other psychophysiological measures such as event-related potential (ERP). Eye tracking and face tracking have been employed in research on partially-automated driving, and systems that measure vigilance and attentional focus are being installed in currently available vehicles, for example as part of General Motors' Supercruise, although, as of this writing, questions remain as to the suitability of such monitoring technology for the task of maintaining driver focus (Bergasa et al., 2006; Lopez, 2019; Wolfe et al., 2020). These mechanisms of operator monitoring present an opportunity for adaptive automation systems to reorient a driver's or operator's cognitive focus to the target task if attention is directed away from it, and to change modality of alerts to increase the likelihood of attending to relevant signals. The screenomics methods of using screenshots to measure task focus in ecological settings are useful for measuring the activity of computer users, but may be highly intrusive and raise the risk of compromising privacy and security (Reeves et al., 2019). The sub-method of measuring switching frequency, rather than evaluating content, may present fewer risks of releasing sensitive information, but is still a surveillance technique. Employing these methods should be considered carefully with respect to the utility of measuring task focus and focus switching, as part of an overall attention to data security and respect for individual user rights, including privacy.

3.3 Adaptable and Adaptive Automation

Adaptable automation provides operators or users the ability to control the level of automated system agency, or potentially to decide if it is active at all. They can therefore tailor the automation's level of agency to their needs, provided they can do so in a wise and safe manner. This freedom can allow operators to address unforeseen circumstances or those beyond automated system capabilities (Miller & Ju, 2015). If Armstrong and Aldrin had been unable to take control of piloting the lunar module when they realized the automated landing system was taking them into a boulder field, they would have had to abort the landing, or could have crashed on the lunar surface. If operators trust the system appropriately and have a well-calibrated mental model of its capabilities, they can use it properly; but if there is a mismatch, adverse consequences may result. It is not always a wise idea to give operators full control of automated systems, especially those meant to ensure safety: the operators of Chernobyl reactor 4 disabled the automatic control rod management system to allow manual control of the reactor beyond its design limits, with catastrophic results (IAEA, 1992). Even though trained, they were inadequately informed and unaware of the potential catastrophic failure modes of the reactor (which were known, but unpublished), and the risks they were taking by pushing the envelope. Had they followed established protocols, or had the controls prevented excursion beyond the reactor's safe operating parameters, there would not have been a disaster. If the system had been designed with an anticipation that operators might attempt to place the reactor in an unsafe configuration, the controls could have been designed to prevent this. Understanding users is therefore a critical part of system design, as it permits the needed flexibility without opening opportunities for abuse (Parasuraman & Riley, 1997).

The flexibility of adaptable automation also can add task load, most critically at times of high stress, or when automated systems may not be engaged in a supervisory role when needed, if they must be manually engaged. Cook et al. (1991) term this problem "clumsy automation" and consider "user-centered automation" to be a solution. User-centered automation takes into account the needs of the users and context, using a technique such as cognitive function analysis (Boy, 1998) to foresee such conflicts. Especially in environments with multiple agents, both technological and human, it is important to determine the optimal design for technologies, the configuration of human and machine roles, and to ensure adequate communication between them. In designing automated systems and their controls, interfaces should encourage proper use and cue recognition of optimal responses, rather than forcing operators, likely under stress, to use imperfect memory or slow checklists. Therefore, the design process should include evaluation of normal and abnormal situation procedures and include how systems can support recovery from adverse events, as well as ways to avoid them.

Adaptive automation, as an extension of adaptable automation (Atkinson et al., 2012; Inagaki, 2003; Kaber, Riley, Tan, & Endsley, 2001) promises to remedy some of the problems of human–system interaction by adapting the actions of the system to address the needs of humans in the moment, due to the level of state, trait, and training factors. Automation can be adaptable through dynamic function allocation: assigning functions to itself, other technological agents, or human operators rather than forcing the task of agency allocation onto humans. This conceptually approaches Licklider's (1960) concept of symbiosis, where there is a close and dynamic relationship between human and machine agents in a partnership arrangement. Designing such a system is similar to adaptable automation systems, with a further focus on how the system should allocate tasks between agents, based upon monitoring the capabilities of a human user or users in the moment, ideally using nonintrusive measures, such as machine vision or assessment of workload.

Adaptive systems may be perceived as behaving erratically or unpredictably if operators are not sufficiently knowledgeable of its dynamic nature and changeable actions. This may reduce

adaptive systems' utility and safety in applications where users cannot be expected to have an accurate mental model and well-calibrated trust model of the system (Atkinson et al., 2012; Fu et al., 2019; Hancock et al., 2011). Research on adaptive automation by Sauer et al. (2013) using a simulated process control task found that with regard to effort expenditure, mental workload, and diagnostic performance, the adaptable automation provided advantages over low and intermediate levels of static automation. Furthermore, operators would select higher levels of automation under conditions of environmental stress compared with times of quiet conditions. While automation has significant benefits, Flemisch et al. (2008) state that automation can also have different drawbacks with respect to the interaction between humans and complex technology. While Parasuraman et al. (2000) proposed two dimensions for the continuous "levels of automation," Flemisch et al. (2008) indicated that the essential elements of automation could be efficiently communicated with a one-dimensional spectrum of continuous automation degrees. Importantly, transitions between levels of automation can be between adjacent levels as well as non-adjacent levels, potentially providing an automation surprise (Sarter et al., 1997) if a system rapidly and dramatically increases or decreases its level of agency. As transitions between levels can be initiated by an operator and by the system, the system must sufficiently communicate changes of state, act in a predictable manner, and prevent inadvertent user-initiated changes of state. Flemisch et al. (2008) further discuss human–machine compatibility in terms of ease of user interaction, user understanding of system actions in each situation, and role-sharing between humans and A^3. This is dependent upon knowing what the A^3 does, why it does what it does, and what its intents will be, creating a mental model of the system's capabilities. The interplay of perceptual, interactive, and intentional processes is illustrated in Figure 7.

Perhaps the earliest adaptive automation systems are the dead-man's switches used in industrial systems and in trains,

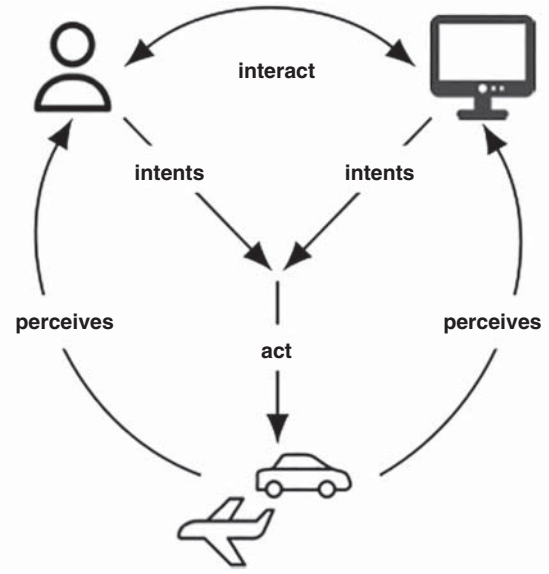

Figure 7 Humans hold a mental model of the system and the environment—and A^3 systems correspondingly hold an image of the human user(s) and environment. The two agents can collaborate in acting on the environment, through operationalizing their intents in an interactive manner. This is a continuous process, where the model of the environment is continually updated through perceptual processes. (*Source:* Adapted from Flemisch et al., 2008.)

which ensure the operator is present and ideally situation aware—and provide an alert or automatically stop the system if not. Vigilance-measuring systems have improved over the past century, and are now becoming a part of the A^3 systems available to members of the general public. Most, if not all, current vehicles with adaptive cruise control and lane keeping (SAE level 2) include some form of vigilance management feature, often a sensor on the steering wheel that detects the presence of a driver's hand. This may ensure that the driver has their hands on the wheel, but does not ensure they are situation aware, attending to the roadway environment, or even awake (Baker, 2019). An approach using eye tracking and/or face tracking can determine if a driver is alert (Caffier, Erdmann, & Ullsperger, 2003), and attending to the ambient environment, but the performance of systems such as General Motors' Supercruise (McGehee et al., 2016) is still insufficient to achieve the aims of ensuring vigilance and situation awareness. As A^3 capacities and sensing abilities increase, it becomes technically feasible to measure and classify eye gaze, and to detect arousal using minimally-intrusive sensors. Ensuring the driver's attention is focused on the roadway is a step in the correct tdirection, but does not encompass all of the needs of vigilance monitoring. Perhaps a positive approach here is ensuring engagement with driving, assisted by automation to reduce active fatigue, rather than fighting passive fatigue due to placing the driver in the role of a largely passive supervisor (Körber, Cingel, Zimmermann, & Bengler, 2015).

Consider a driver who is falling prey to drowsiness, or whose attention is diverted by an alternate task. It would seem logical to require an automated system take over more of the driving task, but this may in fact lead to further disengagement, and reduce the driver's availability if human agency needs to be called upon. Essentially, the design challenge here is to keep the operator in the loop. One solution would be to increase the requirements on the driver, increasing their engagement in the driving task, provided they are capable of driving safely, with help from automation. If the driver cannot drive safely, for example, due to active fatigue or other impairment, the system must take over more of the driving task, and operate without human backup.

3.4 Explorable and Transparent A^3

As previously mentioned, the strategy of documenting A^3 in traditional ways such as regimented training and thick instruction manuals is flawed. A possible remedy is Explorable A^3, designed to be explored without adverse consequences or undue risk. Features analogous to the "undo" capability, or version tracking, are ideal examples of how to allow such exploration, when it is technically feasible. Where actions are irreversible, design should, if possible, support a preview of what an action will do before a user commits to an irrevocable change. For example, an image editing program may provide the option of testing the effect of an image processing filter before it is applied to the source image, while an automated vehicle or a navigation system might show a symbolic simulation indicating that merging into traffic in an upcoming exit would involve a longer delay than driving to the next. Confirmation before an irrevocable action is taken is another example of a strategy from traditional software that can easily be applied in many A^3 applications to avoid costly mistakes, or at least reduce their frequency.

A^3 can provide varying levels of information about intent, performance, future plans, and reasoning processes, and this descriptive characteristic can be referred to as transparency. One way to think about transparency is as providing human teammates with SA regarding purpose, process, performance reasoning process, projected future, and potential limitations (J. Y. Chen et al., 2014). Humans expect to explain their decisions,

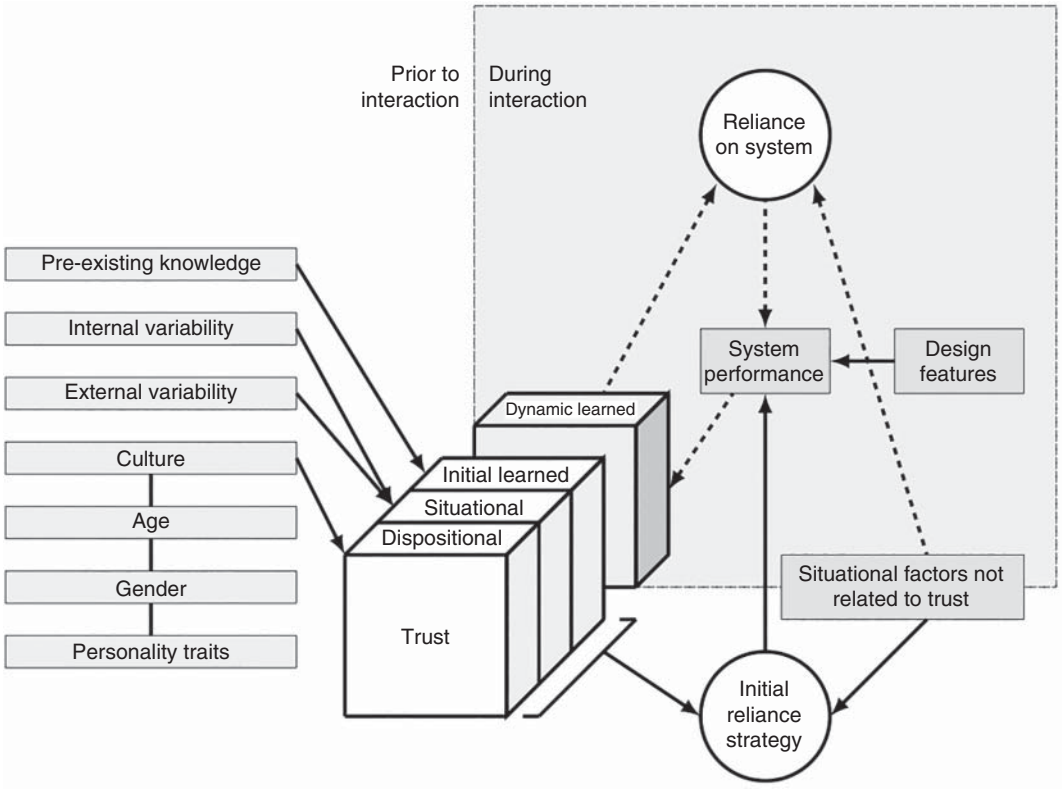

Figure 8 Full model of factors that influence trust in automation and reliance on it, in the categories of personal characteristics, knowledge, and system factors. The dotted arrows represent factors that can change within the course of a single interaction. (*Source:* Adapted from Hoff and Bashir, 2015.)

from trivial to high consequence, to themselves and others, and proponents of Explainable AI suggest holding A^3 to a similar standard. A key factor in this approach is understanding human-to-human communication, especially how humans describe their own reasoning and explain their own choices. This includes acknowledgment of both strengths and weaknesses, discussion of intrinsic biases, and strategic considerations such as available time (Wang & Siau, 2019).

Chen et al. (2014) identified important HF/E issues related to human–A^3 teams in the context of multi-robot control and successful human–agents interactions for the supervision of multiple intelligent systems. These issues include appropriate human trust in the automated systems, individual differences in human–agent (H–A) interaction, efficient human supervision of multiple robots, maintenance of human operator's situation awareness, and retention of human decision authority. They concluded that it is essential to assure automation transparency and include information about human individuals' differences as part of the human–agent design process. Transparency should include the current and future states of multiple intelligent systems, as well as the intent of the agent acting as an intermediate supervisor.

Such capabilities are presently at the edges of technological feasibility, and modifying deep learning techniques to extract explainable features does not necessarily mean that interpretable models will result. As such, most present approaches leverage a combination of such explainable feature extraction and "model induction" in which explanations are inferred from the neural net behavior (Hagras, 2018). In a task, therefore, a system can decide to use available explanations to justify itself to the user. The result, interestingly, is A^3 which cannot always fully explain its own actions, much as one might expect from its human counterpart. An obvious follow-on question is how trust can be appropriately encouraged in users of such systems (and see Hancock et al., 2020).

3.5 Building Trust

Mayer et al. (1995) state that "the *need* for trust arises only in a risky situation" (emphasis in original). While Mayer et al.'s research reviewed trust in human organizations, their conclusions hold in human–A^3 trust as well. This is a result of the way humans unconsciously relate to "interactive media," in a manner similar to their interaction with humans (Nass, Fogg, & Moon, 1996; Nass & Moon, 2000; Reeves & Nass, 1996). Meyer and Lee (2013) describe trust as a cognitive concept, and reliance as a behavioral construct. Conceptually, trust in human–machine interactions encompasses three components: (1) there must be a "truster" to give trust, and there must be a trustee to accept the trust; (2) the trustee must have some sort of incentive to perform the task; and (3) there must be a possibility that the trustee will fail to perform the task. Indeed, these stages mean that trust may be a construct which unfolds over time, both at the individual and societal level (Kaplan et al., 2020).

Hoff and Bashir's trust model includes three layers of variability in human–automation trust: dispositional trust, situational trust, and learned trust (see Figure 8). These components of trust, which in turn influence reliance, are influenced by personal, situational, and system factors. Trust develops over time and is continuously reevaluated with respect to system performance. A system that is initially treated with skepticism can engender trust through demonstration of adequate

performance; a system that demonstrates failures in technical capacity or goal incongruence will erode human trust in it, and as a result reduce reliance on it. Hoff and Bashir (2015), Verberne et al. (2012, 2015), Takayama et al. (2009), and Nass and Moon (2000) recommend that designers should increase the automated system's degree of anthropomorphism, politeness, ease of use, and transparency to promote better calibrated trust in systems and to forestall automation disuse. These social factors related to trust and reliance stand in addition to evaluations of pure technical performance. Verberne et al. (2015) found that an embodied agent system that shows a visible facial likeness to the user can engender greater feelings of trust, independent of behavioral factors. This is analogous to homophily, the way humans instinctively relate more strongly to similar persons (Rogers & Bhowmik, 1970).

Weitz et al. (2020) used A^3-based virtual agents and developed an explainable A^3 system to improve the trust of end users in human–A^3 interaction. Their research concluded that using virtual agents in an explainable artificial intelligence system can improve users' trust. Also, the study indicated that combining voice output with the virtual agent was more effective than using virtual agents alone. Siau and Wang (2018) identified that trust is crucial in developing and accepting A^3 technology. The main elements of trust in automation are shown in Figure 8, while the technology features of A^3 that affect trust-building are illustrated in Figure 9 and Figure 10.

Rapid advances in A^3 have created new opportunities for users to interact with automation (Oh et al., 2018). Van Maanen et al. (2005) indicated that a vast proliferation of A^3 applications leads to a shift from simple human–machine interactions to

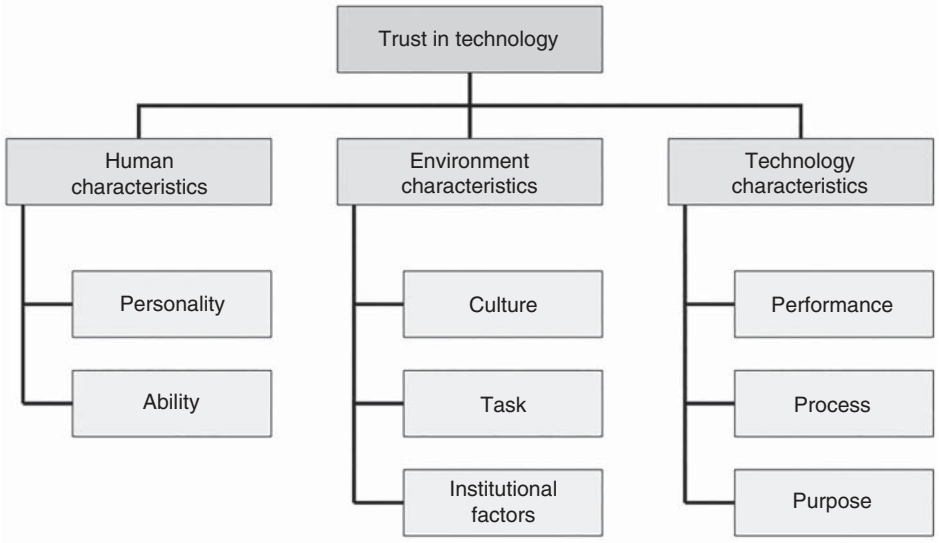

Figure 9 Factors affecting trust in technology include human user characteristics, environmental factors, and features of the technology. (*Source:* Adapted from Siau and Wang, 2018.)

Initial trust formation	Continuous trust development
Performance	Performance
• Representation • Image/perception • Reviews from other users	• Usability and reliability • Collaboration and communication • Sociability and bonding • Security and privacy protection • Interpretability
Process	Process
• Transparency and ability to explain • Trialability	• Job replacement • Goal congruence

Figure 10 Trust in automation is formed by assessment of performance and the system's communicative abilities. It is sustained by, and further developed or retarded by, continuous evaluation of performance and considerations regarding goal congruence and worries over the place of humans in the human-automation environment. (*Source:* Adapted from Siau and Wang, 2018.)

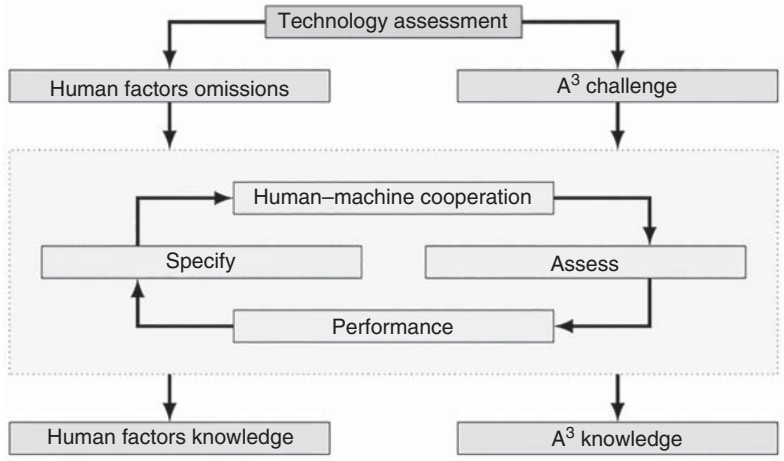

Figure 11 To discover gaps in both human factors and A^{3}knowledge with respect to human-machine teams, a technology assessment focused on human factors and A^{3}limits and capacities in cooperation was conducted by van Maanen et al. 2005. (*Source:* Based on van Maanen et al., 2005.)

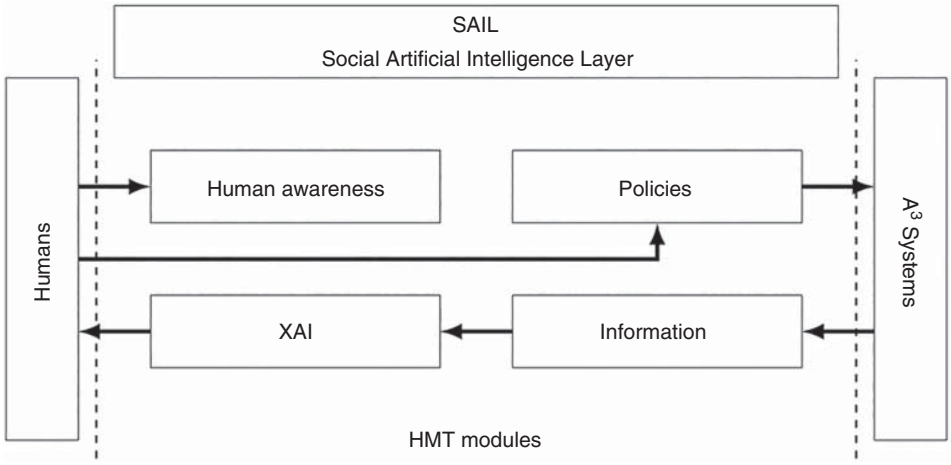

Figure 12 In human–machine teams, a social artificial intelligence layer mediates the communication between humans and A^3 systems. This element includes modules for policy setting determined by human users, and information flow from the A^3 system, through A^3 to human operators. (*Source:* Adapted from van Maanen et al., 2005.)

more complex cooperative human–A^3 interactions that require a multidisciplinary approach. The development process of such human and A^3 cooperative interaction is shown in Figure 11.

Sujan et al. (2019) discussed using A^3 in patient care and the necessity to put much more effort into developing the new A^3-based technologies. They advocate for increased focus on human factors in A^3 technology for clinical processes and services, with due consideration of automation bias, impact on human performance, handover, situation awareness, and patient interaction. Van der Vecht et al. (2018) discussed the promising paradigm of human–machine teaming (HMT) and introduced a framework for the efficient development of HMT concepts. The proposed framework has a modular social layer between autonomous systems and human team members to improve teamwork as shown in Figure 12.

A^3 systems need to be both trustworthy, and communicate that they are trustworthy, in order to encourage users to form appropriate trust in them (J. D. Lee & See, 2004). A risk is overtrust, and systems which appear to have greater capacities than they actually exhibit can engender inappropriate reliance, leading to adverse situations. As in human–human trust, the results of A^3 overtrust are diverse, and sometimes strongly negative. A related threat is de-skilling: if operators rely on a system that is normally trustworthy and helpful such as the glide-slope indicator or auto-landing system in an aircraft, problems can result when that system is unavailable to help. This over-reliance has been noted as a contributing factor in several airplane crashes where pilots' skills were insufficient without the assistance of automation.

3.6 Security Design

It is comforting to imagine a world where all humans and A^3 together work to conquer outside challenges, rather than one where some portion of humans and machines *are* those challenges, but no design is complete without considering security. Human users represent the largest attack surface in most socio-technical systems (Sebescen & Vitak, 2017). Malicious

cyberthreat actors are acutely aware of this vulnerability and deliberately work to exploit it by employing Remote Online Social Engineering (ROSE) techniques to influence individuals inside the organizations they target. Phishing, and vishing (voice phishing), have become household terms, daily occurrences, and a constant source of threat.

While algorithmic cybersecurity works valiantly to eliminate the threat of ROSE attacks, A^3 to augment cybersecurity are presently not the most effective form of protection. HF/E practitioners have long argued that design should be leveraged as a deterrent, and a defensive boon, but these efforts are still all too rare (Gutzwiller, 2015; Vieane et al., 2016). In fact, the primary method that organizations use to counter ROSE is training, and the design and implementation of security education, training, and awareness (SETA) programs for employees (Y. Chen, Ramamurthy, & Wen, 2015; D'Arcy & Hovav, 2008; Yoo, Sanders, & Cerveny, 2018). A characteristic shortcoming of this training is that it often provides the same training to all employees, regardless of job role, technical sophistication, or personal susceptibility to ROSE attacks. A major component of SETA programs is the employment of simulated phishing email campaigns that mimic actual malicious phishing email campaigns. These simulated phishing emails are intended to provide a form of inoculation for employees against malicious emails. This training is very realistic and can be one of the most effective ways to prepare users to avoid real phishing attacks as the campaigns replicate attacks that have successfully and adversely affected other organizations (Carella, Kotsoev, & Truta, 2017).

Simulated phishing also provides valuable data on employee responses to simulated phishing attacks, which allows researchers to better understand the characteristics of user responses. From this data, two interesting archetypes of user patterns begin to emerge. First are a subgroup termed "repeat clickers." These individuals are those employees who continually fall victim to phishing attacks, regardless of message content or environmental influences (Canham, Posey, & Bockelman, 2020). A 2015 PhishMe report states that in an analysis of over 8 million simulated phishing emails "67% of employees who respond to simulated phishing attacks are repeat victims and therefore likely to respond to phishing emails more than once" (PhishMe, 2015). Thus, while some in the general population might fall for phishing attacks occasionally, this subgroup is much higher than the average. The second group represents the other tail of the distribution. These individuals actively seek to determine whether an email is phishing and then report those emails they believe to be malicious, to help protect the organization and their fellow colleagues from those threats. These users are labeled "protective stewards" whose actions help assist rather than detract from organizational cybersecurity efforts (Burns et al., 2018; Posey et al., 2013). When they believe they have detected deception, employees have the option of ignoring the email or reporting the email as a suspected attack. When reporting, some organizational security departments ask users to forward the emails to their security representatives so that they can inspect the message for possible phishing activity, and others have activated technological solutions that make it easier for user reports of suspected phishing attacks to take place. Often, the inclusion of report buttons within email interfaces greatly simplifies this task by quarantining the email, alerting the security staff of the potential threat, and releasing the email back to the employee if the security staff deems it a false alarm. This action of making security easier for users represents a best practice that should be deployed more widely.

Resources within information security departments are often severely limited. Being able to better allocate the limited resources that are available would provide significant benefit to security operations staff. Research suggests that resources should not be uniformly distributed but rather should consider the relative strengths and weaknesses of an organization's user population (Canham et al., 2020). While security departments should tailor their SETA efforts toward their respective employee subpopulations, designers can also contribute significantly to ensuring security against threat actors by building systems that will enable security operations staff to tailor solutions to different users according to their individual susceptibility patterns rather than applying the same levels of security protocol regardless of individual differences. Adding additional protections for the most vulnerable users, while empowering those who show the most promise toward protecting the organization, security departments will be able to increase their own effectiveness and see a better return on investment with already limited resources. Likewise, SETA programming strongly centered on phishing would be more beneficial for the employees who are repeat clickers rather than forcing those who are protective stewards to endure the same intervention. Such an approach indicates the need for organizational SETA programs to go beyond decreasing risk by also incorporating aspects that help build, strengthen, and maintain employee motivations to engage in positive behaviors.

3.7 Design Strategies and Frameworks

In human-centered design of A^3 systems, important strategic differences exist between approaches to focusing upon the human. For example, constrained design and unconstrained design are engineering-based approaches to addressing the increased demand that accompanies human–A^3 interaction, especially in contexts where a high-stakes activity involves a crucial primary task, such as driving (Skrypchuk, Langdon, Sawyer, & Clarkson, 2020). Constrained design approaches provide goals for the maximum portion of user capacity an automated system should occupy, and may in fact be codified within an organization, legislated, or otherwise related to standards. An example of constrained design would be the limiting of typing input functionality while a vehicle is in motion. Because this approach fundamentally curtails the overall capabilities of the system, and in doing so curtails the available actions available to a user, constrained design can result in frustrated users. Motivated, frustrated users may indeed actively work to circumvent behavior intended as a protection. This can lead to a "cat and mouse" situation, and in extreme cases to systems with significant risks that sit alongside those risks they were intended to prevent. For example, legislatively mandated alcohol interlock devices installed on the vehicles of individuals with a history of drunk driving have grown so complex that they are themselves a significant distracted driving hazard. Software interlocks on current generation cell phones require a multiple click effort to disable while driving, potentially leading to a similar pitfall. Conversely, unconstrained design considers trade-offs between interaction with A^3 and users, and works to balance interaction design with competing goals. An example can be found in guardian angel automated driving approaches, which assist the driver in driving safely, and should the driver be unable to safely accomplish the driving task, automation can guide the vehicle out of the roadway and park safely at the side of the road. It is important to realize that these design perspectives are not exclusionary of one another, and that, given the overriding goal of task success in safety, either and both may be appropriate. It is also important to realize that whenever user action is curtailed, there is the potential for loss of trust and attempts to circumvent limitations (Guttman & Gesser-Edelsburg, 2011).

The Design Thinking process (Gibbons, 2016) is one framework which can be applied, in concert with consideration of HF/E principles, to increase the likelihood of producing usable,

safe, and pleasurable products of design. Design begins with user research: who will use the product, service, or system, and to pursue what goals? What risks are there of misuse or error? Observing the current users in the contexts where use is expected will provide an important base upon which development can begin. It is essential to empathize with users and understand what they do and under what conditions. What do they say about the current situation? What do they feel about it? The second stage is to define problems that can be addressed. Designers must develop insights from the empathetic research stage, by identifying unmet needs, desires, and pain points. Designers then ideate, brainstorming a large range of ideas; the more, the better. Involving users in the ideation stage can be helpful, especially if specialized knowledge is necessary for the tasks performed. Prototypes can vary in fidelity from simple sketches to full simulations--and for this process, all are valuable. Cheap and fast prototypes are ideal for concept refinement, high fidelity prototypes can be used for user testing before committing to production. Testing can be time-consuming and expensive, but with A^3, implementation can be enormously costly, and errors can have deadly or society-damaging consequences. The final step is implementation, and ongoing evaluation of the performance of the product. Especially with software or software-defined systems, updates can be applied to remedy deficiencies. Every step in this process should refer back to prior stages, and work in harmony with the empathetic research stage. Constant reference to the needs of intended users will help to produce a product that meets needs and satisfies actual wants, rather than assumed ones (see Figure 13).

Value Sensitive Design (VSD) (Friedman, 1996; Friedman, Kahn, & Borning, 2002; Friedman, Kahn, Borning, & Huldtgren, 2013) provides a toolset for incorporating human values into the design process, and it is well worth considering the moral and ethical values that are inevitably embedded into products of design. VSD is comprised of three investigations: *conceptual*, *empirical*, and *technical*. The conceptual component is comprised of a stakeholder analysis, a determination of who is influenced by or interact with a system, as well as a determination of stakeholders' values and an analysis of the tensions between values. An example of stakeholders related to an anesthesia delivery system would include physicians and nurses, as well as patients, even if patients never interact with the system's interface, and the values they would pursue would include safety, efficiency, and system flexibility. The empirical component encompasses research that informs designers' understanding of user needs, situated behaviors, and practices. Technical investigation covers analysis of the affordances of the technologies employed, focusing on artifacts rather than stakeholders. These three components, integrated into a user-centered or participatory design process, can aid in surfacing and addressing the values of stakeholders and the tensions between them. An example of values in tension is the desire to provide privacy and security of personal data, and the need to provide enormous training datasets to A^3 systems.

Designers must weigh the risks of capturing and storing sensitive data which could be exposed, against the benefits anticipated from the deployment of a system that could perform important work, such as identifying markers of disease. They must also constantly consider the contribution, or lack thereof, of the human in-the-loop. Radiologists partnering with detection A^3 may miss cancers, and drivers partnering with ADAS may miss roadway obstacles, and these failures among others can be minimized through design consideration.

3.8 Tandem Failure and Mutual Reinforcement

If an automated system is supervised by a human, or a human is supervised by a technological system, and both fail in their

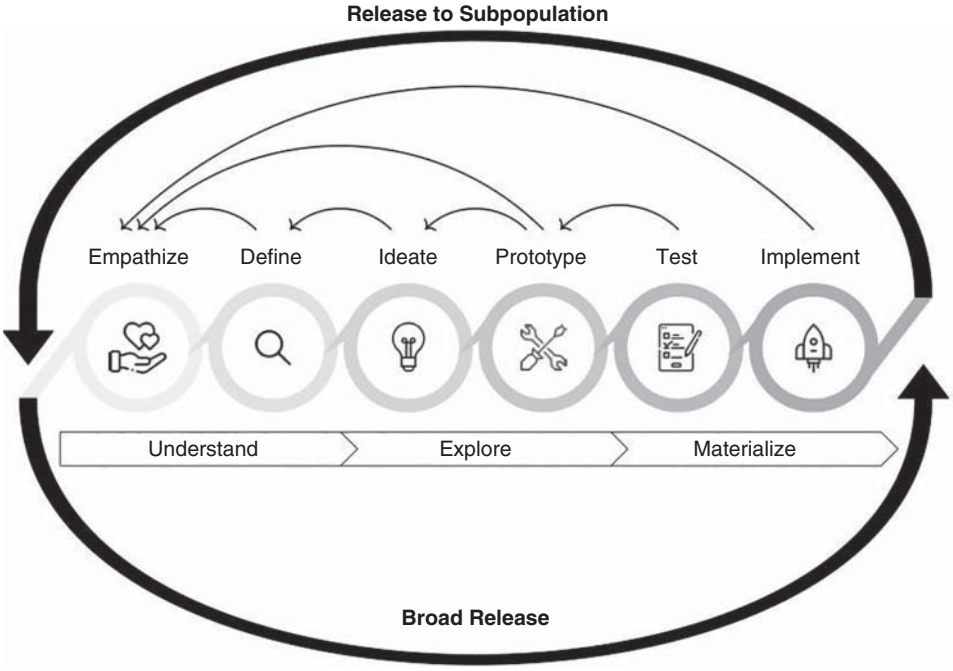

Figure 13 Frameworks for the design carried out by humans can be incorporated into the ongoing improvement of A^3 systems, and that improvement itself can be the work of hybrid intelligence. An iterative version of an A^3 technology can be analyzed to build understanding, explored, and new capabilities materialized before being released to a subpopulation to gather data for an additional cycle, while at the same time feeding into the broad release of the A^3 product. As such, teams of humans and A^3 together refine an overarching A^3 technology. (*Source:* Adapted from Gibbons, 2016.)

Table 1 Humans and A^3 Can Act as Mutually-Reinforcing Agents. If Both Fail at the Same Time, It May Be Difficult or Impossible to Recover from a Tandem Failure Situation

		A^3	
		Correct	Incorrect
Human	Correct	All is well	Operator should override system
	Incorrect	System should override operator	

tasks at the same time, the combined system will fail to achieve success in the overall task (Sheridan, 1992a). The stakes have been as high as "apocalyptic." We are reading this only because in the early hours of September 26, 1983, Stanislav Petrov evaluated the signals from the Soviet Union's early warning satellite network against his own knowledge and intuition and deemed the alert of a NATO attack erroneous (Lebedev, 2011). We all owe a debt of gratitude not only to this courageous individual, but to those who trained him, and the designers who built a system capable of conveying subtext allowing him to deduce that the attack he saw was false. Automation has prevented disasters as well, both large and small. The lives saved by antilock brakes and electronic stability control go largely unsung, but these are but a few of the technologies preventing automotive crashes that otherwise would have occurred (Farmer, 2004, 2006). Many A^3 successes in other areas go likewise uncounted—improving medical diagnostics and weather analysis improves the life of millions, but this is often overshadowed by stories of high-profile failure.

In many applications, humans monitor systems, providing an error-trapping role. Unfortunately, humans are ill suited to supervision, especially prolonged supervision (N. H. Mackworth, 1948; Sheridan, 1992a). Fitts et al. (1951) state that systems should monitor human performance, which places the near-infinite attention of a machine in a role to which it is well suited, and can allow humans to use their skills in applications demanding cognitive flexibility and evaluation. Combining human and machine capacities to reinforce each other offers the greatest total envelope of performance and safety, and roles should be assigned based upon an understanding of human and machine capabilities, both in terms of trait and state factors. Adaptive automation (Inagaki, 2003; Kaber et al., 2001) can assist in this, by allowing for dynamic change in roles, dependent upon conditions. Unfortunately, there is no way to completely solve the problem of tandem failure (see Table 1) where both human and A^3 systems fail simultaneously, but good system design can reduce the prevalence and seriousness of those situations. They can also be mitigated through testing, and training, ideally in a low-risk proxy environment.

3.9 Testing and Training with Simulation

For A^3 systems designers, simulation is an opportunity to obtain vital data inexpensively, and without real-world risk. For human operators, simulation provides a similar opportunity to acquire experience interacting with A^3 systems, and is invaluable to pilots, drivers, physicians, and others. Simulation can be used for research and for training simultaneously, and HF/E research regularly involves simulated environments in which hybrid teams react to real-world challenges. Indeed, a major use of simulation is for research and regulatory validation of human–A^3 interaction design. With the increasing availability and decreasing barriers to use of simulation, simulation is sure to be a growth area for data generation, human training, and industrial, academic, and regulatory research.

At its best, simulation offers high experimental and ecological validity (Campbell, 1957; Orne & Holland, 1968; Reichardt, 2011), ability to elicit naturalistic behavior (Harrison, Haruvy, & Rutström, 2011), and high applicability for psychology, HF/E, and human–computer interaction research (Blascovich et al., 2002). Virtual reality, either using a head-mounted display or CAVE-type environment (Cruz-Neira, Sandin, & DeFanti, 1993), which includes automotive and flight simulators, provides an ideal environment for both research and training, provided they offer a sufficient level of presence (IJsselsteijn et al., 2000; K. M. Lee, 2004a; Sheridan, 1992b; Slater & Usoh, 1993). Presence can be described as a sense of "being there" (Igroup Project Consortium, 2015) in a virtual environment, and high presence elicits naturalistic behavior as a result of cognitive processes related to human treatment of media experiences as if they were real experiences (K. M. Lee, 2004b; Reeves & Nass, 1996). Even if we know that the simulator cannot move, and the other vehicles on the "road" are merely images, drivers in the simulated environment act as if the virtual threats are real, both physiologically and behaviorally.

"Fidelity" is a common term in simulation, too often linked to visual complexity and too little linked to its more useful grounding in relevance to the question being asked. Some of the earliest transportation simulators, Link trainers (Jeon, 2015) which reproduced the instrument readings and simulated proprioceptive cues of flight to provide instrument flight training, can be described as high fidelity, and produced valid re-creations useful to human trainees with high efficiency and no crash risk. In the following years, simulators became more advanced, and integrated increasingly sophisticated visual features, first through cameras on gantries over miniature environments, and later via computer-generated scenes. While high mundane realism (Aronson, Wilson, & Brewer, 1998; Difonzo, Hantula, & Bordia, 1998; Dobbins, Lane, & Steiner, 1988) is ideal, photorealistic visuals are not necessary. Rather, providing as engaging an environment as possible is more important in terms of eliciting naturalistic responses. This scene-setting can include a narrative around the experiment, potentially including a degree of deception, to avoid cueing participants to consciously control behavior to match experimenter expectancies. Validity can also be enhanced by removing or minimizing edges of the virtual environment so that the real world minimally intrudes, and by using time pressure (if warranted by the situation) to focus attention on the issue at hand and reduce participant ability to think about the situation at a meta-level before responding. Breaking presence by stopping time, for example, when using the SAGAT methodology to study situation awareness (Endsley, 1988; Sirkin, Martelaro, Johns, & Ju, 2017) or when studying decision making where social desirability bias is a significant validity threat (Wintersberger, Frison, Riener, & Hasirlioglu, 2017), may reduce the ecological validity of findings, and thus choices regarding adjusting time in simulation research should be carefully evaluated. Therefore, fidelity and ecological validity together must be considered to determine whether a simulator is providing valid data, for whatever purpose.

A drawback of simulation for A^3 training and validation is that the environment does not have the richness of the outside world, and so human trainees may be ignorant of true risks and machines trained on the data may be fragile, exhibiting edge

case behavior in naturalistic environments. This may present a problem for testing A^3 systems in simulated environments, as has been employed in the autonomous vehicle space with the A^3 software being tested in simulated environments rather than on the road (Huang et al., 2016). Constructed environments can strategically over-represent features that are rare in the natural environment (Kelly et al., 2018), an important benefit of simulation when compared with testing and validation natural environments. The threats of the real world present an open set, where in addition to polite and predictable pedestrians, there are all manner of unpredictable things out in the road environment. Animals, fallen cyclists, unpredictable human behavior, obstacles in the roadway, all must be properly included in simulations if the A^3 driving systems they train are to avoid edge case behavior. Even if the risk of a fatal crash is between 4-40 per billion vehicle-kilometers, there are many vehicle-kilometers driven every day, and thus over a million fatalities worldwide annually.

Simulation provides an ideal locale for training (Kozak, Frison, Riener, & Hasirlioglu, 1993) and has long been used for training users in how to use automated systems. This has in the past been limited to professional users such as pilots, military personnel, and industrial operators. Today, the rapidly improving capabilities of VR, increasing ease of constructing virtual environments, and dramatically decreasing price, have all made simulation a broadly viable training mechanism. While a full-motion simulator or complete nuclear power plant control suite is an ideal environment, even a low-cost HMD or desktop VR system can provide adequate presence for training, and for research, and with the proliferation of such systems enabled by the continued increase in computing power available for economical prices, this is a strong growth area with many opportunities available to provide training where it was previously difficult or impossible.

Simulation is notably useful for studying driver–vehicle interaction, as simulators can reveal driver behavior without physical risk, and with a perfectly repeatable driving environment, which is not possible on the road. Most drivers have minimal formal training and thus automated systems must be designed to offer adaptive support that can accommodate new drivers, older and impaired drivers, and experts, through user control or self-adjustment. To ensure such systems work well for a wide variety of drivers, substantial usability research is needed, and this can be afforded through simulation of A^3 systems in environments and situations where reliance on or conflict with A^3 will occur. The future is a big place, and so simulation is moving far beyond driving, or any single context. Indeed, a substantial number of human factors researchers and engineers are now engaged in simulating potential systems-level futures for A^3–human teamwork, and determining whether they are worthy of investment, or avoidance.

4 TOWARD, AND BEYOND, A *PRIMUM NON NOCERE* OF A^3

At this moment in history, humanity's brightest minds are hard at work attempting to create at least one new type of intelligence through A^3. The discussion of what this intelligence should be, how it should interact with humans, and whether we should trust it, is in no way new. A^3 has been part of human storytelling and philosophy for thousands of years, automata emerging as playthings of the Greek gods, the mythical Golem, machines like the Mechanical Turk, and modern-day mass internet movements of humans collectively building illusions of autonomy (see also Geoghegan, 2020). In very narrow contexts, this dream is already realized, and we stand on the precipice of having it realized far more broadly. Hancock describes this remapping as a rising tide of increasing A^3 capability, at first eroding the major continents of human activity, then eventually relegating select activities to increasingly isolated and shrinking islands. In this world, what were for the most part philosophical debates have become pressing business, regulatory, and personal considerations. In this conversation regarding what machines should be to humans, HF/E is vital, not only to human-centered A^3, but to the quickly evolving system element, driven by continued cross-pollination of ideas and technologies. Here, we briefly ground the reader in the history that has brought us to these questions. We then argue that human-centered realizations of A^3, a new intelligence to join humans, must in fact be human-centered. We introduce the reader to HF/E concepts that may underpin a desirable future, one in which we have designed A^3 to provide much benefice, but, and first, do no harm.

Mechanical systems, like the Jacquard loom, Babbage's Difference Engine, Hollerith's tabulators, and mechanical computers for gun-laying, presented the first real instances of automation, replacing human thought with "programmed" deterministic devices. Human ingenuity was thus separated temporally and physically from operation: a programmer punched the cards or ground the cams at one time and place, and later those determined the pattern woven into the cloth or the elevation of the guns. This separation meant that decisions had to be made ahead of time, but for these relatively mechanistic processes the considerations were limited in scope. By the 1970s and 1980s, electronic computerized devices were widespread, replacing mechanical systems in vehicles, machine tools, and even toys. However, these systems did similar work to their earlier counterparts. For example, electronic fuel injection is more flexible and offers superior performance compared with a carburetor, but fills the same function. While the systems available to the public were for the most part analogous to their earlier counterparts, sophisticated semi-autonomous systems were introduced into military, medical, aviation, and industrial settings. These had both benefits and costs. For example, the Aegis combat system improved air defense and fleet management, but also was in part responsible for the tragic erroneous shootdown of an Iranian civilian airliner, mistaken for a military aircraft potentially vectoring to attack a US Navy ship (Kopec & Tamang, 2007).

Increasing technical capabilities due to advances in computer technology led automation to expand into the public sphere, with highly complex or non-deterministic systems appearing in home devices, such as thermostats, microwaves, and other appliances. The idea that a microwave might be an acceptable embodied agent would have been the premise for a joke only years ago, but we are now surrounded by such highly agentic systems (D. Miller, 2016). These can be enormously beneficial, and quite frustrating, perhaps at the same time. These so-called cyber-physical systems (CPS) can reduce energy use, help us improve our health, conveniently deliver what we want, and provide access to the contents of the world's information stores. They also change the nature of human agency, exert persuasive power over us (Fogg, 2003), and can elicit enormous frustration when they do not operate the way users expect or desire them to (Pernice, 2015), or if there is a mismatch in goals (Verberne et al., 2012).

Humans have extended their cognitive capacities through external mechanisms since the advent of drawing and writing thousands of years ago, and continue to do so with low-tech devices like pen and paper, and increasingly agentic devices such as computers and networks. The first edition of this volume, *The Handbook of Human Factors and Ergonomics* was published in 1987. It seems so long ago, when computers were not yet in nearly every pocket, and finding the references cited meant a trip to a library to find another book or journal. Now

the many papers cited are only seconds away, and intelligent agents can recommend similar sources to go with the ones referenced by the author. As this chapter is being written, active work is being conducted on more efficient recommender systems, augmented reality, wearable computing devices, and brain–computer interfaces. These technologies further blur or erase the line of where human cognition ends and where the "outside" begins. The mind of an author is expanded by access to sources and the ability to store information not in memory, but on paper or in electronic media. When knowledge is even less intermediated than it is now, perhaps by a direct neural connection to the internet, what will that mean in terms of defining humanity and A^3? We have already turned visionary dreams like Vannevar Bush's conceptual Memex (Shneiderman, 1998) and formerly science-fiction concepts like Penny's wrist-top computer from the cartoon *Inspector Gadget* into readily available products. Will the visions of Masamune Shirow's brain–computer interfaces be that far off, and what limits will there be on new technologies that further blur the line between the natural and the technological?

Presently, there is more work than ever to be done. While the history of human factors and ergonomics is closely tied to automation, the connection between these disciplines and machine learning-based technologies is more recent and tenuous. A search of Google Scholar on '"machine learning" AND "human factors" OR "ergonomics"' yielded 1,830 returns in the year 2000, 12,200 returns in 2008, and 68,000 returns in 2020. This is a relatively low number as compared to 2020 results of 2,950,000 for the term "machine learning" and 2,440,000 for "human factors" OR "ergonomics." Related terms show similar patterns, which reveal a relatively small but growing intersection in the literature between the two fields. It has been noted that educational intersections, meanwhile, are lagging, both in terms of individuals educated in machine learning but versed in HF/E principles, and the converse (Hannon et al., 2019). Given the complex problems outlined in this chapter, there are significant benefits to be realized in the change in education, culture, and resource allocation that is and will be necessary to continue, and indeed accelerate, the union of the HF/E with machine learning communities.

4.1 Future Challenges in A^3 Design

"First, do no harm" is a mandate humanity has given to its physicians, but some of its physicians will now be not human, but A^3. There are many technical challenges in order for intelligent automation to reach the level of maturity that is required to ensure effective collaboration with humans (Abbass, 2019), but arguably just as many design challenges. Klien et al. (2004) identify ten such challenges for human–AI collaboration (see Table 2), and these continue to be pressing issues in 2020. These challenges encompass the topics of regard for human agency, trust and trustworthiness, transparency, goal alignment, and bidirectional communication and awareness of the partner's state. In many contexts, we need to consider how to work with A^3 toward the goal of doing "no harm."

There are many technical challenges for intelligent automation to reach the level of maturity that is required to ensure effective collaboration with humans (Abbass, 2019), but arguably just as many design challenges. Klien et al. (2004) identify ten such challenges for human-A^3 collaboration, and these continue to be pressing issues in 2020 (see Table 2). These challenges encompass the topics of regard for human agency, trust and trustworthiness, transparency, goal alignment, and bidirectional communication and awareness of the partner's state.

Parasuraman et al. (2000) identify four stages of information processing and action as relevant to automation: information acquisition, information analysis, decision selection, and action implementation. Abbass (2019) identifies a set of risks stemming from the interaction of human control and agent control, broken down by Parasuraman et al.'s stages (see Table 3). The nature of these risks vary both as a function of overall division of authority, and the stages of the decision–action sequence as assigned to a human or technological agent.

Table 2 Ten Challenges for Intelligent Agents and Human-Agent Teams

Challenge 1: To be a team player, an intelligent agent must fulfil the requirements of a Basic Compact (a commitment of goal alignment) to engage in common-grounding activities.
Challenge 2: To be an effective team player, intelligent agents must be able to adequately model the other participants' intentions and actions vis-à-vis the joint activity's state and evolution - for example, are they having trouble? Are they on a standard path proceeding smoothly? What impasses have arisen? How have others adapted to disruptions to the plan?
Challenge 3: Human-agent team members must be mutually predictable.
Challenge 4: Agents must be directable.
Challenge 5: Agents must be able to make pertinent aspects of their status and intentions obvious to their teammates.
Challenge 6: Agents must be able to observe and interpret pertinent signals of status and intentions.
Challenge 7: Agents must be able to engage in goal negotiation.
Challenge 8: Support technologies for planning and autonomy must enable a collaborative approach.
Challenge 9: Agents must be able to participate in managing attention.
Challenge 10: All team members must help control the costs of coordinated activity.

Source: Adapted from Klien et al., 2004.

These risks are in part compounded and in part mitigated by the inherent sociality of agents, even ones that do not appear to be humanlike. The truly social nature of the relationship between humans and A^3, described by Reeves and Nass (1996) as "interactive media" follows similar patterns to the relationships between humans. We unconsciously treat computers as if they are humanlike entities (Nass & Moon, 2000), treating them with politeness, trusting them in the same ways (Atkinson et al., 2012), and being persuaded by them (Fogg, 2003; Fogg & Nass, 1997). This should be kept in mind in terms of understanding how people relate to systems, for example, developing fanciful mental models of or constructing personalities for machines (Barley, 1988), or treating robots as irreplaceable and awarding "fallen" robotic comrades who have no true agency military medals (Garber, 2013). As humanlike agents, automata should be considered in a psychological frame, as well as an engineering one, especially ones that exhibit more social features, such as speech, faces, and intentionally, and unintentionally, social behaviors.

4.2 A^3 and Machine Ethics

Bryson and Winfield (2017) defined intelligence as the capacity to do the right thing at the right time. Coincidentally, Stanislav Petrov, the Soviet military officer who wisely overruled a faulty automated system, described himself as "just in the right place, at the right time" in his acceptance speech for the World Citizen Award in 2006 (Anthony, 2013). For the foreseeable future,

Table 3 Cooperative Human-Machine Control Introduces Risks that Are a Function of the Locus of Sense-Making, Decision-Making, Execution Ability, and Execution Authority

Human control	Sense-making	Decision-making	Execution ability	Execution authority	Nature of risk
Absolute	H	H	H	H	Limited human cognition and bounded rationality could lead to high errors, information overload, and inability to manage complex tasks.
High	A^3	H	H	H	Undesirably biased analytics could drive the human to unfair decisions, while human bias and limited cognition could add more complexity to the mix.
High	H	A^3	H	H	Undesirably biased recommendations could make the human accountable for unethical or legally uncompliant decisions, although the human could be overwhelmed by the available data, and their own bias and limited cognition could add more complexity to the mix.
Medium	A^3	A^3	H	H	In the absence of transparency and explainability of the A^3, the human does not have enough information to form a judgement regarding the chosen decision. Information and situation complexity could overload the human. The human could become accountable for inappropriate decisions.
Low	A^3	A^3	A^3	H	In the absence of transparency and explainability of the AI, the human has no understanding of the rationale of the decision. Information and situation complexity could overload the human. The human's accountability is blinded.
Low	A^3	A^3	H	A^3	The AI controls human actions and could lead the human to wrong actions.
None	A^3	A^3	A^3	A^3	The human is out of the loop, legal responsibilities and accountabilities regarding the decision are both unclear.

Source: Adapted from Abbass, 2019.

human-A^3 teams will be needed in critical roles, to ensure optimal performance and safety, with both agents acting to prevent error on the part of the other.

Winfield et al. (2019) note that that in the near future, autonomous systems will need to make decisions that have ethical consequences. Machine ethics can be classified into in four major categories (1) Ethical Impact Agents: any machine that can be evaluated for its ethical consequences, (2) Implicit Ethical Agents: Machines that are designed to avoid unethical outcomes, (3) Explicit Ethical Agents: Machines that can reason about ethics, (4) Full Ethical Agents: Machines that can make explicit moral judgments and justify them. The relevant societal and regulatory implications of machine ethics, including the question of ethical governance, have also been discussed by many, both from the philosophical and psychological standpoint, and from the standpoint of the technologies in play. The field of ethical AI, ethical robots, and machine ethics, focuses on the question of how A^3 can behave ethically from both philosophical and engineering standpoints (Chatila & Havens, 2019). These fundamental questions are (1) "should society delegate moral responsibility to its machines?," and (2) "how to build an ethical machine?" Winfield et al. (2019) also discussed how to test artificial agents, advancing the concept of an "Ethical Turing Test" which would compare the choices of an artificial moral agent with those of humans. This presents its own quandary, specifically, should artificial moral agents (Floridi & Sanders, 2004) be designed to be similar to humans, or should they espouse different morals?

Nebeker et al. (2019) argue that it is essential to determine who is responsible for advancing the ethical practices of A^3 technology and development that addresses the gaps and provides recommendations in the area of ethical principles of A^3 applications. These are considered under the overarching ethical principles of (1) respect for persons, that people are given appropriate autonomy to make choices; (2) beneficence, defined as benefits outweighing potential risks and harms; and (3) justice, that inappropriate burdens are not placed on individuals (see Figure 14). The sub-concepts stemming from these include privacy, access and usability, appropriate data management, and considerations of relative risks and benefits.

Dignum (2018) distinguished three primary levels of A^3 ethics, including (1) Ethics by Design, i.e., considering different elements of human factors in human–A^3 interactions; (2) Ethics in Design, i.e., implementing different regulatory and engineering methods to support the ethical implications of A^3; and (3) Ethics for Design, which encompasses codes of conduct and standards for ethical practice. Yu et al. (2018) propose a taxonomy for A^3 governance with four areas of consideration: (1) exploring ethical dilemmas; (2) individual ethical decision frameworks; (3) collective ethical decision frameworks; and (4) ethics in human–A^3 interactions. Jobin et al. (2019) further identify a need for a global agreement on principles and guidelines for ethical A^3, including six fundamental ethical principles of (1) transparency; (2) justice; (3) fairness; (4) non-maleficence; (5) responsibility; and (6) respect for privacy. They also point out the importance of integrating guidelines and development efforts in A^3 with substantive ethical analysis and adequate implementation strategies.

Recently, IEEE's Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems released a discussion document called Ethically Aligned Design (How, 2018). The presented report includes general principles of how to embed values into autonomous intelligent systems. This includes methods to guide ethical design, safety, and beneficence of Artificial General Intelligence (AGI) and artificial superintelligence, personal data and individual access control, and discusses how to reframe autonomous weapons systems, economics and humanitarian issues, and law (Chatila & Havens, 2019).

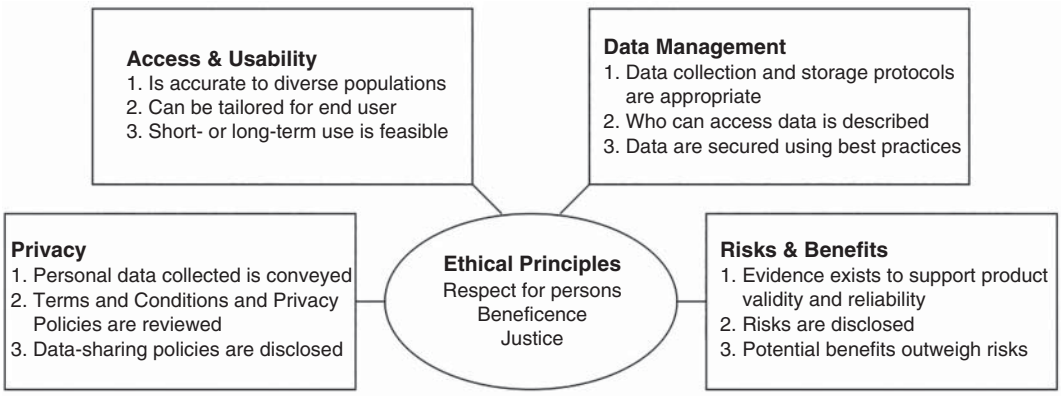

Figure 14 Ethical issues associated with A^3 include respect for privacy, access to and usability of systems, appropriate security and access control, and a balance of risks and benefits. These fall under the overarching principles of respect for persons, beneficence, and justice. (*Source:* Adapted from Nebeker et al., 2019.)

The Association for Computing Machines has developed a code of ethics and professional conduct (ACM Code 2018 Task Force, 2018), as has the Human Factors and Ergonomics Society (Human Factors and Ergonomics Society, 2020). Four IEEE working standards groups (Shahriari & Shahriari, 2017) are developing candidate standards to address related ethical concerns, including the following: P7000—Model Process for Addressing Ethical Concerns During System Design; P7001—Transparency of Autonomous Systems; P7002—Data Privacy Process; and P7003—Algorithmic Bias Considerations. These can serve as a starting point for considering professional obligations regarding work in human factors or computer engineering, and both stress the obligation to the public good. To that end, it is incumbent upon engineers, designers, and academics to avoid work that crosses ethical boundaries and unduly increases the risks of harm, for example, by offering an opportunity for discrimination or introduction of bias, especially that which causes further disadvantage to the most vulnerable.

Following legal and professional guidelines and endeavoring to avoid doing harm is a fair first level of concern for design, but there is a further onus to actively do good. This is not necessarily straightforward, and there are substantial differences in moral and ethical alignments, due to personal factors (Graham et al., 2013), cultural factors (Awad et al., 2020; Huang et al., 2016), and contextual forces. It is impossible to provide a prescriptive statement other than that it is imperative for designers and engineers to consider the moral and ethical consequences of their choices that become embedded in systems. Without deliberate attention to algorithmic bias (Kirkpatrick, 2016), systems can easily reproduce extant patterns of favor and disfavor, as has been noted in study of recommender systems determining criminal recidivism risk (Khademi & Honavar, 2019). To quote Angela Y. Davis, "In a racist society, it is not enough to be non-racist, we must be anti-racist." Ruha Benjamin (2019) calls out "Jim Code" - a continuing devaluation of Blackness and the result of avoiding addressing issues of race in the design of A^3 systems, resulting in continuing structural racism. A^3 systems that purport to increase efficiency may instead increase inequality, unless design and policy decisions are made to ensure that these adverse effects do not result (Eubanks, 2017). It is therefore the responsibility of educators, researchers, designers of technologies, and practitioners to take up the call to push back against an unjust status quo. To ignore these social dimensions of HF/E work is to fail to make the future a better place, especially for those who are most disadvantaged.

A^3 systems almost inevitably embed our heuristics and biases. Machine learning systems encode the parameters of the training set, which may reflect societal biases such as the overrepresentation of Black Americans ensnared in the judicial system (Angwin, Larson, Mattu, & Kirchner, 2016), or underrepresentation of minorities in medical research (Gianfrancesco, Tamang, Yazdany, & Schmajuk, 2018). Algorithms need to be audited to determine if there are latent biases in the system, even if biases are not explicitly coded into them (Diakopoulos, 2016). Considering the risks of incarceration, medical error, deepening wealth inequality, or other adverse outcomes, care must be taken with algorithmic systems to discover potential harms (Peña Gangadharan, Eubanks, & Barocas, 2014) and deliberately de-bias algorithmic systems to prevent such eventualities. Virginia Eubanks, in *Automating Inequality* (2017), describes how technology is often used to surveil and police the disadvantaged. While this is a policy choice, A^3 makes it easier to track behavior and detect any deviation from the system's rules, and thus deny benefits as a result. Eubanks cites multiple examples of how burdens are disproportionately placed on the most vulnerable, as a result of automation failure (for example, in records processing), and through the actualization of policy decisions. For a designer or engineer, the charge is then not solely to deliver technical excellence, reducing failure of systems that could result in social harms, but also to consider the implications of building systems that can cause disproportionate social harm, not only in cases of failure, but that cause harm by design.

Biases in A^3, as it currently exists, are not rare (Nissenbaum, 1996), but there are design strategies to play fair. Zimmerman and Di Rosa (2019) identify two types of fairness relevant to algorithmic bias: procedural fairness, where impartial rules and the same kind of data are equally applied to all; and substantive fairness where the outcomes and their societal impacts are considered with respect to justice. There are places for both of these: fairness in processing should avoid situations such as poorly trained image processing algorithms misclassifying Black humans as gorillas (Vincent, 2018), or which overstate the risk posed by a potential parolee, dependent upon racially and socially linked factors that do not reflect their actual level of recidivism risk (Angwin et al., 2016). Ensuring substantive fairness is more difficult, and politically more sensitive. A^3 can magnify the existing biases in contemporary society, or can be used to reduce them, but this requires deliberate action on the part of designers. A^3 should, where possible, provide parity for the disadvantaged, and address the needs of those

who are marginalized. Value-sensitive design (Friedman, 1996) can determine the needs of stakeholders, and tools such as A^3 auditing (Raji et al., 2020) can begin movement toward a more fair and equitable use of A^3 to champion pro-social goals, avoid hazards, or simply not anger end users.

4.3 Moral and Ethical Values in Tension

To whom does a technological agent owe a moral duty? Is a user or owner owed special consideration, or is an agent's duty to the general good? Should an A^3 toy owe a duty to a child, and keep a secret if asked to, or should it report child abuse if confided in, or tell others of dangerous behavior (Jones & Meurer, 2016)? Should one's vehicle protect the occupants over all others? Daimler released a statement to the press stating just that, but quickly walked it back (Taylor, 2016). These questions that seem academic at the moment soon will not be so: systems need to be programmed to act (or reject action) for situations that will occur to someone, somewhere, at a future time and place.

A significant amount of philosophical debate, as applied to automated vehicles, has been conducted (Bonnefon, Shariff, & Rahwan, 2015; Lin, 2017), as well as survey research (Awad et al., 2018, 2020; Bonnefon, Shariff,& Rahwa,n 2016; Li et al., 2016; Malle et al., 2015) and research in virtual reality (Navarrete et al., 2012; Sütfeld, Gast, König, & Pipa, 2017; Wintersberger et al., 2017). This knowledge can inform design of vehicle systems as well as shape the development of other agents ranging from shopping agents to military systems. In addition to determining what to program systems to do in situations of moral dilemma, a second question is, how much human autonomy to allow in such situations? In some contexts, where time and cognitive resources are available, human input can be considered and put to use in shaping outcomes. Under time pressure and uncertainty (Maule et al., 2000; Svenson & Maule, 1993), human cognition and action can be compromised, and thus it may be necessary to curtail human autonomy in such contexts, especially if the risks of injury or damage are significant. This presents a philosophical dilemma, a question of how values of autonomy and rationality are to be valued, especially when designing a system that can take agency away from a human, potentially to seek their considered goals, or to execute predetermined actions when human action is not possible or unwise in the moment.

An example of how technology is used to resolve these conflicts are the many self-binding technologies: they allow the rational self to force the situated self to follow a predetermined course of action (Fogg, 2003; Thaler & Sunstein, 2009). A time-lock box can help one focus by locking away a phone or snack, which is then inaccessible until the timer runs out. A more critical example is the design of lane keeping systems with mechanisms to avoid overcorrection and potential loss of control (Boink et al., 2014).

4.4 Bad Actors and Dark Patterns

It is worth considering that transparency is not always the goal of A^3, and that some automation exists to assist users toward actions or processes they do not want to engage in, and which may indeed harm them. Effective user interface design guides user behavior, and this can occur with or without the user's conscious awareness. Ideally, A^3 design can indeed guide the user along a course of action that is beneficial to them, under their own direction or in pursuit of their overarching goals. In contrast, user interface design techniques that guide users along courses of action that are not in their best interest are collectively referred to as Dark Patterns (Brignull, 2020). Waltzman (2017) provides several categories of dark patterns: these comprise (1) Nagging: Redirection of expected functionality that persists beyond a few interactions; (2) Obstruction: Making a process more difficult than it needs to be with the intention of dissuading certain actions; (3) Sneaking: Attempting to disguise information that is relevant to the user; (4) Interface Interference: Manipulation of the interface that privileges certain actions over others; and (5) Forced Action: Requiring users to perform certain actions in order to maintain access to a service. Each has implications for trust, but each use by designers to infringe upon user trust also has implications for the culture and lineage of design to come. Our path toward the A^3 of the future will be paved with the design decisions of the present.

Lacey and Caudwell (2019) describe three additional characteristics of dark patterns: (1) dark patterns produce the illusion of user sovereignty; (2) they emphasize short-term gains over long-term benefits; and (3) they manipulate, manage, and exploit emotion in order to generate "data myopia" (Stark, 2016) in the user. Most studies to date have focused on dark design patterns in online environments, but at least one group of researchers have examined these patterns as applied to cyber-physical systems in the form of home robots (Caudwell & Lacey, 2020). Home assistant robots have access to the most private aspects of our domestic lives, and many collect and transmit significant amounts of data to their servers. From a design standpoint, potential security concerns for end users can, be answered through frameworks intended to address cyber-attacks, for example, the Three Pillars of Information Security. These include (1) confidentiality: preventing unauthorized disclosure of information; (2) integrity: preventing unauthorized modification of information; and (3) availability: maintaining access to information (Merkow & Breithaupt, 2014; Smith, 1989). Taken together, the pillars provide a useful framework when considering security concerns in general, and if a system design violates any of these pillars, designers are advised to consider them carefully and signal this to end users appropriately.

A^3 design can be incredibly difficult to design in terms of walking the fine line between responsive behavior toward users, and violations of their trust and rights. It is very important to avoid knee-jerk labeling of such violations as malicious, as the very nature of autonomy means that such actions may be unintended, or inadvertent. We, here, call out that fine line between necessity and malfeasance, and also acknowledge that it can be quite subtle when building A^3 technologies. It is certainly a design consideration, and one that touches upon practically every previous section of this chapter, among many areas of concern. Intentional dark patterns and unintentional behavior that look similar, can both have the same outcomes: violations of user rights, expectations, or trust. Of course, true dark patterns must be designed, and so a related concern is how designers decide to engage their talents. Unlike physicians, our constellation of professions has no ethical consensus to do no harm. It is worth considering at the outset of system architecture whether transparency will be the goal of your A^3, and how justice and beneficence will be integrated into the design process.

4.5 You Are What You Engineer

Perhaps the most definable exceptional human capacity is our ability to incorporate our tools into ourselves. This is not merely a philosophical statement, as modern neuroscience has shown clearly that, as they are learned, tools such as chisels and violinists' bows acquire brain regions that are deeply analogous to biological components, such as arms or fingers (Nicolelis, 2011; Ramachandran, 1998, 2012). For an organism so committed to incorporation of outside components, it is perhaps unexceptional that humans have long endeavored to literally integrate our tools. In a more contemporary sense, we have come incredibly far over the last decade, and is it indeed now

possible to refer to the "early neuro-hacking days." Linking cochlear implants directly to a digital cassette output (Gfeller & Lansing, 1991), or Wes Warwick connecting his human arm to a robotic one over the early Internet (Warwick, 2016) seem almost quaint. They have now been joined by recent, but assuredly soon-to-seem antiquated, efforts by technology companies to directly connect human brains to smart devices. Industry groups like Facebook and Open Water are working to advance noninvasive technologies able to read from, and write to the human nervous system and brain, leveraging near infrared and holographic techniques for monitoring neural blood flow patterns in real time (Openwater, 2020). Simultaneously, Neuralink, Kernel, and others are working on invasive strategies to measure and influence the electrical activity of the brain with implanted electrodes, those being connected to computing devices allowing "read and write" capabilities.

Many lines are blurred as the human nervous system is being increasingly linked to technological systems, and while it is unquestionable that these technologies hold great promise to augment human capabilities, there will simultaneously be many perils. Augmenting humans through such technology, both cognitively and physiologically, is unquestionably a task best suited to A^3. A^3 possess the required flexibility, and could meet neuroplasticity "halfway" while integrating with the most complex computational device humans possess: the brain and peripheral nervous system. It was one week before this writing that Neuralink demonstrated pigs with Neuralink brain implants, and they made the claim of intending to start human trials within the year. The stated goal of this company is to provide an avenue through which to integrate human cognition with that of A^3. In this, however, there are some additional considerations that designers, and indeed everyone else, should consider.

We may well be witnessing a prelude to the end of the era in which human thought is generally accepted to be an entirely private activity. This thought may be odd to many readers, but the authors have great confidence it will be less so to our children. Indeed, readers born after 1975 or so will have grappled with their parents' discomfort with the "death of privacy" ushered in by the modern internet, and as in that case the outcomes do not have to be negative. The horizon for determining this is not far off. Significant progress has been made in recent years in the development of both invasive and non-invasive brain–machine interfaces (BMIs), allowing operators to communicate directly with machinery (computers, robotics, cars, artificial limbs, etc.) using only their thoughts (Roelfsema, Denys, & Klink, 2018). A Google patent search reveals that over 400 patents were filed for brain-connected devices in 2019 (Google, 2020). The success of either of these technologies, practical external neuroimaging or computer-connected "permanent" electroencephalography, will open a new universe of possibilities for the realm of neurosecurity. This intimate connection between a user's neural system and technological space opens an entirely new dimension of attack surfaces that may potentially be exploited by cyber threat actors, or by actors one would hope to be more benevolent. As in the internet parallel above, conflicts of interest are certain to abound. Should Neuralink succeed in producing technologies that aid in decision-making, how should a Neuralink implant assist a user regarding a decision about upgrading to the new model?

Existing frameworks are appropriate for considering the coming questions of neurosecurity (Canham & Sawyer, 2019). Within the context of neurosecurity, breaching the pillar of confidentiality would potentially allow unprecedented access to an individual's most private data: their unfiltered thoughts. A breach of integrity would mean that an attacker could inject commands into a neuro-device, or alternatively send false signals to the brain directly through the device. A failure of availability would prevent a user from being able to control the device or receive data from it. While these failures may seem like pure science fiction, proof of concept attacks have already been demonstrated for each pillar (Roelfsema et al., 2018). These attack surfaces apply equally to both the connected devices themselves, but also to the human nervous system. An attacker with access to a connected device may utilize this access to passively monitor activity and gather information, but might also use this access to manipulate the behavior of the connected human. Consider the case of cochlear implants which augment human hearing capability. These hearing augmentation devices might be compromised and used as mobile audio surveillance devices, creating "human bugs;" however, these compromised devices are also capable of producing painful stimuli. Leveraging this capability, a threat actor could expose a subject to positive or negative stimulation in response to specific actions and thereby manipulate behavior. Imagine coupling painful stimulation with geo-fencing to keep a user confined within a specified boundary, in a way that could not be avoided or stopped. Considering the significant risks of direct neural interface augmentation, security should not be left as an afterthought but rather, integrated into the design from earliest possible phase. That earliest possible time is right now.

5 ENGINEERING A^3 EVERYONE LOVES

Automation, autonomy, and artificial intelligence (A^3) technologies serve as extensions to human ability and may soon rewrite the way that humans work, live, and even think. The authors are hopeful that humanity will like, and even love, what is built (Egwatu, Sawyer, & Hancock, 2019; Hancock, 2014; Hancock, Pepe, & Murphy, 2005; Nyholm & Frank, 2017; Nyholm & Smids, 2020). It is likely that the present generation of computational tools, algorithms that can learn from data, are not the end point for tools which enable technologies that act in reasonable, and even humanlike ways. As ubiquitous as it seems now, machine learning may soon be a curious historical footnote. What will remain is the art and science of designing intelligent teammates, which is not only the focus of this chapter, but also perhaps the most consequential intellectual endeavor of our moment in human history. The desire for synthetic intelligent creations has been a staple of human desire for so long, it is indeed exciting to be alive at the moment when these ambitions are coming to fruition. While artificial general intelligence (AGI) remains, at present, just a dream, a number of promising, and promised, advances give at least some possibility that younger readers of this chapter will live to meet them. Of course, when Simon predicted machines "capable … of doing any work a man can do" (Chase & Simon, 1973), he had similar sentiments, and today we may be no different. We do not need to live to meet this technology in order to help decide what it will be: the design decisions we make will inform the philosophies and technologies that are its building blocks. In developing present-day, highly useful A^3 technologies, knowledge from human factors and ergonomics (HF/E) is of great use, especially to designers with the difficult task of dovetailing humans and machines in complex systems. Technology serves as a greater extension of human ability each year, and optimal performance still comes from hybrid teams, something that will arguably be true even after we meet, or become one with, AGI. We must work in the present to design machines capable of accompanying us into the sometimes chaotic environments to come, and helping us to enjoy that journey.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Katherine Rahill for assistance with the initial literature review, Sarah Minion, for assistance with figures, Esat Boucaud for assistance with citations

and citation management, Adam Beal and Lexi Neame for influential conversations, and Marian W. Sawyer for assistance with proofreading.

REFERENCES

Abbass, H. A. (2019). Social integration of artificial intelligence: Functions, automation allocation logic and human-autonomy trust. *Cognitive Computation, 11*(2), 159–171. https://doi.org/10.1007/s12559-018-9619-0

ACM Code 2018 Task Force. (2018, June 22). *ACM code of ethics and professional conduct*. Association for Computing Machinery. https://www.acm.org/code-of-ethics

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy, 47*, 1–6. https://doi.org/10.1016/j.infoecopol.2019.05.001

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=e_IIJeE3oY9zIykeZdUxxAunfmrN8x1-

Anthony, P. (2013, April 27). *The Man Who Saved the World*. Statement Film. http://themanwhosavedtheworldmovie.com/

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G.Lindzey (Eds.), *The handbook of social psychology* (Vols. 1–2, 4th ed., pp. 99–142). New York: McGraw-Hill.

Atkinson, D., Hancock, P., Hoffman, R. R., Lee, J. D., Rovira, E., Stokes, C., & Wagner, A. R. (2012). Trust in computers and robots: The uses and boundaries of the analogy to interpersonal trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*(1), 303–307. https://doi.org/10/gfkwbb

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1911517117

Bainbridge, L. (1983). Ironies of automation. In *Analysis, design and evaluation of man-machine systems* (pp. 129–135). Amsterdam: Elsevier.

Baker, P. C. (2019, November 27). 'I think this guy is, like, passed out in his Tesla.' *The New York Times*. https://www.nytimes.com/2019/11/27/magazine/tesla-autopilot-sleeping.html

Barley, S. R. (1988). The social construction of a machine: Ritual, superstition, magical thinking and other pragmatic responses to running a CT scanner. In M. Lock & D. Gordon (Eds.), *Biomedicine examined* (pp. 497–539). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-2725-4_19

Behymer, K. J., & Flach, J. M. (2016). From autonomous systems to sociotechnical systems: Designing effective collaborations. *She Ji: The Journal of Design, Economics, and Innovation, 2*(2), 105–114.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Cambridge: Polity.

Ben Zur, H., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica*, *47*(2), 89–104. https://doi.org/10/d3bxbg

Bergasa, L. M., Nuevo, J., Sotelo, M. A., Barea, R., & Lopez, M. E. (2006). Real-time system for monitoring driver vigilance. *IEEE Transactions on Intelligent Transportation Systems, 7*(1), 63–77. https://doi.org/10/ft9jpk

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry, 13*(2), 103–124. https://doi.org/10.1207/S15327965PLI1302_01

Boink, R., van Paassen, M. M., Mulder, M., & Abbink, D. A. (2014). Understanding and reducing conflicts between driver and haptic shared control. In *2014 IEEE International Conference On Systems, Man and Cybernetics (SMC)* (pp. 1510–1515). http://ieeexplore.ieee.org/abstract/document/6974130/

Bolton, C., Machová, V., Kovacova, M., & Valaskova, K. (2018). The power of human–machine collaboration: Artificial intelligence, business automation, and the smart economy. *Economics, Management, and Financial Markets, 13*(4), 51–56.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2015). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? *ArXiv:1510.03346 [Cs]*. http://arxiv.org/abs/1510.03346

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10/f8rqvz

Boy, G. A. (1998). Cognitive function analysis for human-centered automation of safety-critical systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '98* (pp. 265–272). https://doi.org/10.1145/274644.274682

Bradshaw, J. M., Hoffman, R. R., Johnson, M., & Woods, D. D. (2013). The seven deadly myths of "autonomous systems". *IEEE Intelligent Systems, 28*(3), 54–61.

Brignull, H. (2020). Dark patterns. https://darkpatterns.org/

Broadbent, D. E. (1958). Effect of noise on an "intellectual" task. *The Journal of the Acoustical Society of America, 30*(9), 824–827.

Bruckner, L. A. (1978). On Chernoff faces. In *Graphical representation of multivariate data* (pp. 93–121). Amsterdam: Elsevier.

Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer, 50*(5), 116–119.

Burns, A. J., Roberts, T. L., Posey, C., Bennett, R. J., & Courtney, J. F. (2018). Intentions to comply versus intentions to protect: A VIE theory approach to understanding the influence of insiders' awareness of organizational SETA efforts. *Decision Sciences, 49*(6), 1187–1228. https://doi.org/10.1111/deci.12304

Cades, D. M., Crump, C., Lester, B. D., & Young, D. (2017). Driver distraction and advanced vehicle assistive systems (ADAS): Investigating effects on driver behavior. In N. Stanton (Ed.), *Advances in human aspects of transportation* (pp. 1015–1022). Berlin: Springer.

Caffier, P. P., Erdmann, U., & Ullsperger, P. (2003). Experimental evaluation of eye-blink parameters as a drowsiness measure. *European Journal of Applied Physiology, 89*(3–4), 319–325. https://doi.org/10.1007/s00421-003-0807-5

Caldwell, B. S., Megan, N.-Y., & Jordan, R. (2019). Advances in human-automation collaboration, coordination and dynamic function allocation. *Advances in Transdisciplinary Engineering, 10*, 348.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297–312. https://doi.org/10.1037/h0040950

Canham, M., Posey, C., & Bockelman, P. S. (2020). Confronting information security's elephant: The unintentional insider threat. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented cognition. Human cognition and behavior* (pp. 316–334). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-50439-7_22

Canham, M., & Sawyer, B. D. (2019). Neurosecurity: Human brain electro-optical-signals as MASINT. *American Intelligence Journal, 36*(2), 41–47.

Canonico, L. B., Flathmann, C., & McNeese, N. (2019). Collectively intelligent teams: Integrating team cognition, collective intelligence, and AI for future teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*(1), 1466–1470.

Carella, A., Kotsoev, M., & Truta, T. M. (2017). Impact of security awareness training on phishing click-through rates. *2017*

IEEE International Conference on Big Data (Big Data) (pp. 4458–4466). https://doi.org/10.1109/BigData.2017.8258485

Caudwell, C., & Lacey, C. (2020). What do home robots want? The ambivalent power of cuteness in robotic relationships. *Convergence, 26*(4), 956–968. https://doi.org/10.1177/1354856519837792

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55–81. https://doi.org/10/c226rf

Chatila, R., & Havens, J. C. (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* (pp. 11–16). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_2

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency*. Army research lab: Aberdeen proving ground md human research and engineering.

Chen, Y., Ramamurthy, K. (Ram), & Wen, K.-W. (2015). Impacts of comprehensive information security programs on information security culture. *Journal of Computer Information Systems, 55*(3), 11–19. https://doi.org/10.1080/08874417.2015.11645767

Chollet, M., Ochs, M., & Pelachaud, C. (2017). A methodology for the automatic extraction and generation of non-verbal signals sequences conveying interpersonal attitudes. *IEEE Transactions on Affective Computing*.

Chung, S. (2019). Email usage: Working age knowledge workers. 2019 Adobe Email Study. https://www.slideshare.net/adobe/2019-adobe-email-usage-study

Clough, B. T. (2002). Metrics, schmetrics! How the heck do you determine a UAV's autonomy anyway? Wright-Patterson AFB, OH: Air Force Research Lab.

Coelingh, E., Eidehall, A., & Bengtsson, M. (2010). Collision warning with full auto brake and pedestrian detection—A practical example of automatic emergency braking. *2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 155–160. https://doi.org/10/bgfvv2

Coleman, F. (2019). *A human algorithm: How artificial intelligence is redefining who we are*. Berkeley, CA: Counterpoint Press.

Cook, R. I., Woods, D. D., Mccolligan, E., & Howie, M. B. (1991, January 1). Cognitive consequences of clumsy automation on high workload, high consequence human performance. Paper presented at Fourth Annual Workshop on Space Operations Applications and Research (SOAR 90), Lyndon B. Johnson Space Center, TX. https://ntrs.nasa.gov/citations/19910011398

Cruz-Neira, C., Sandin, D. J., & DeFanti, T. A. (1993). Surround-screen projection-based virtual reality: the design and implementation of the CAVE. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 135–142).

Cuevas, H. M., Fiore, S. M., Caldwell, B. S., & Strater, L. A. (2007). Augmenting team cognition in human-automation teams performing in complex operational environments. *Aviation, Space, and Environmental Medicine, 78*(5), B63–B70.

D'Arcy, J., & Hovav, A. (2008). Does one size fit all? Examining the differential effects of is security countermeasures. *Journal of Business Ethics, 89*(1), 59. https://doi.org/10.1007/s10551-008-9909-7

Dash, R., McMurtrey, M., Rebman, C., & Kar, U. K. (2019). Application of artificial intelligence in automation of supply chain management. *Journal of Strategic Innovation and Sustainability, 14*(3).

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. JSTOR. https://doi.org/10.2307/249008

De Melo, C. M., Carnevale, P., Read, S., Antos, D., & Gratch, J. (2012). Bayesian model of the social effects of emotion in decision-making in multiagent systems. Paper presented at Conference on Autonomous Agents and Multi-agent Systems and 55–62.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56–62. https://doi.org/10.1145/2844110

Difonzo, N., Hantula, D. A., & Bordia, P. (1998). Microworlds for experimental research: Having your (control and collection) cake, and realism too. *Behavior Research Methods, Instruments, & Computers, 30*(2), 278–286. https://doi.org/10.3758/BF03200656

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1–3. https://doi.org/10.1007/s10676-018-9450-z

Dobbins, G. H., Lane, I. M., & Steiner, D. D. (1988). A note on the role of laboratory methodologies in applied behavioural research: Don't throw out the baby with the bath water. *Journal of Organizational Behavior, 9*(3), 281–286. https://doi.org/10.1002/job.4030090308

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 45–70.

Drew, T., Võ, M. L.-H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science, 24*(9), 1848–1853. https://doi.org/10.1177/0956797613479386

Egwatu, C., Sawyer, B. D., & Hancock, P. A. (2019). Perspectives: Digital influences on sexual discourse in disabled populations. *Critical Disability Discourses/Discours Critiques dans le Champ du Handicap, 9*(0), Article 0. https://cdd.journals.yorku.ca/index.php/cdd/article/view/39748

Eisma, Y. B., Hancock, P. A., & de Winter, J. C. (2020). On Senders's models of visual sampling behavior. *Human Factors*, 0018720820959956.

Endsley, M. R. (1987). The application of human factors to the development of expert systems for advanced cockpits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 31*(12), 1388–1392.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National* (vol. 3, pp. 789–795). https://doi.org/10.1109/NAECON.1988.195097

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32–64. https://doi.org/10/ftd9tz

Endsley, M. R. (2016). *Designing for situation awareness: An approach to user-centered design*. Boca Raton, FL: CRC Press.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(2), 381–394. https://doi.org/10/bj6f3c

Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor* (1st ed.). New York: St. Martin's Press.

Farmer, C. M. (2004). Effect of electronic stability control on automobile crash risk. *Traffic Injury Prevention, 5*(4), 317–325. https://doi.org/10/b8bcph

Farmer, C. M. (2006). Effects of electronic stability control: An update. *Traffic Injury Prevention, 7*(4), 319–324. https://doi.org/10/d836zp

Fereidunian, A., Lucas, C., Lesani, H., Lehtonen, M., & Nordman, M. (2007). Challenges in implementation of human-automation interaction models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 1–6.

Fitts, P. M., Viteles, M. S., Barr, N. L., Brimhall, D. R., Finch, G., Gardner, E., Grether, W. F., Kellum, W. E., & Stevens, S. S. (1951). Human engineering for an effective air-navigation and traffic-control system. DTIC Document.

Flemisch, F., Schieben, A., Kelsch, J., & Löper, C. (2008). Automation spectrum, inner/outer compatibility and other potentially useful human factors concepts for assistance and automation. *Human Factors for Assistance and Automation*.

Floridi, L. (2020). AI and its new winter: From myths to realities. *Philosophy & Technology, 33*(1), 1–3. https://doi.org/10.1007/s13347-020-00396-6

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

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

Fogg, B. J., & Nass, C. (1997). Silicon sycophants: The effects of computers that flatter. *International Journal of Human-Computer Studies, 46*(5), 551–561. https://doi.org/10/cz2mgc

Frazier, M. L., Johnson, P. D., & Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *Journal of Trust Research, 3*(2), 76–97. https://doi.org/10.1080/21515581.2013.820026

Friedman, B. (1996). Value-sensitive design. *Interactions, 3*(6), 16–23. https://doi.org/10/dntfwj

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington Technical Report*, 02–12.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). value sensitive design and information systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laboratory* (pp. 55–95). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7844-3_4

Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge: Cambridge University Press.

Fu, E., Sibi, S., Miller, D., Johns, M., Mok, B., Fischer, M., & Sirkin, D. (2019). The car that cried wolf: Driver responses to missing, perfectly performing, and oversensitive collision avoidance systems. *2019 IEEE Intelligent Vehicles Symposium (IV)*, 1830–1836. https://doi.org/10.1109/IVS.2019.8814190

Garber, M. (2013, September 20). Funerals for fallen robots. *The Atlantic*. https://www.theatlantic.com/technology/archive/2013/09/funerals-for-fallen-robots/279861/

Gelderblom, H., & Menge, L. (2018). The invisible gorilla revisited: Using eye tracking to investigate inattentional blindness in interface design. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces* (pp. 1–9). https://doi.org/10.1145/3206505.3206550

Geoghegan, B. D. (2020). Orientalism and informatics: Alterity from the chess-playing Turk to Amazon's Mechanical Turk. *Ex-Position, 43*, 45–90.

Gfeller, K., & Lansing, C. R. (1991). Melodic, rhythmic, and timbral perception of adult cochlear implant users. *Journal of Speech, Language, and Hearing Research, 34*(4), 916–920.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine, 178*(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Gibbons, S. (2016, July 31). *Design thinking 101*. Nielsen Norman Group. https://www.nngroup.com/articles/design-thinking/

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton-Century Crofts.

Google. (2020, September 3). Google patents: Brain computer interface. Google Patents. https://patents.google.com/?q=%22brain+computer+interface%22&after=priority:20190101

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Academic Press. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Greenlee, E. T., DeLucia, P. R., & Newton, D. C. (2018). Driver vigilance in automated vehicles: Hazard detection failures are a matter of time. *Human Factors, 60*(4), 465–476.

Griggs, T., & Wakabayashi, D. (2018, March 20). How a self-driving uber killed a pedestrian in Arizona. *The New York Times*. https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html

Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. *Interaction Studies, 8*(3), 483–500. https://doi.org/10/gfkv9m

Guttman, N., & Gesser-Edelsburg, A. (2011). "The little squealer" or "the virtual guardian angel"? Young drivers' and their parents' perspective on using a driver monitoring technology and its implications for parent-young driver communication. *Journal of Safety Research, 42*(1), 51–59.

Gutzwiller, R. S., Fugate, S., Sawyer, B. D., & Hancock, P. (2015). The human factors of cyber network defense. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2016, 59*(1), 322–326.

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer, 51*(9), 28–36.

Halasz, F. G., & Moran, T. P. (1983). Mental models and problem solving in using a calculator. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '83* (pp. 212–216). https://doi.org/10.1145/800045.801613

Hancock, P. A. (2014). Automation: How much is too much? *Ergonomics, 57*(3), 449–454. https://doi.org/10.1080/00140139.2013.816375

Hancock, P. A. (2018). On the design of time. *Ergonomics in Design, 26*(2), 4–9. https://doi.org/10.1177/1064804617735018

Hancock, P. A. (2020a). Driving into the future. *Frontiers in Psychology, 11*, 2405.

Hancock, P. A. (2020b). The humanity of humanless systems. *Ergonomics in Design, 28*(3), 4–6.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., Visser, E. J. de, & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2020). Evolving trust in robots: Specification through sequential and comparative meta-analyses. *Human Factors*, 0018720820922080

Hancock, P. A., Pepe, A. A., & Murphy, L. L. (2005). Hedonomics: The power of positive and pleasurable ergonomics. *Ergonomics in Design*, *13*(1), 8–14.

Hancock, P. A, Sawyer, B. D., & Stafford, S. (2015). The effects of display size on performance. *Ergonomics, 58*(3), 337–354.

Hannon, D., Rantanen, E., Sawyer, B. D., Ptucha, R., Hughes, A., Darveau, K., & Lee, J. D. (2019). A human factors engineering education perspective on data science, machine learning and automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*(1), 488–492.

Harris, W. C., Hancock, P. A., Arthur, E. J., & Caird, J. K. (1995). Performance, workload, and fatigue changes associated with automation. *The International Journal of Aviation Psychology, 5*(2), 169–185. https://doi.org/10.1207/s15327108ijap0502_3

Harrison, G. W., Haruvy, E., & Rutström, E. E. (2011). Remarks on virtual world and virtual reality experiments. *Southern Economic Journal, 78*(1), 87–94. https://doi.org/10.4284/0038-4038-78.1.87

Heer, J. (2019). Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences, 116*(6), 1844–1850.

Hoc, J.-M. (2000). From human–machine interaction to human–machine cooperation. *Ergonomics, 43*(7), 833–843. https://doi.org/10.1080/001401300409044

Hockey, G. R. J., & Wiethoff, M. (1993). Cognitive fatigue in complex decision-making. In S. L. Bonting (Ed.), *Advances in space biology and medicine* (Vol. 3, pp. 139–150). Amsterdam: Elsevier. https://doi.org/10.1016/S1569-2574(08)60101-X

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.

Hollnagel, E., & Bye, A. (2000). Principles for modelling function allocation. *International Journal of Human-Computer Studies, 52*, 253–265. https://doi.org/10.1006/ijhc.1999.0288

Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors, 48*(1), 196–205. https://doi.org/10.1518/001872006776412135

How, J. P. (2018). Ethically aligned design. *IEEE Control Systems Magazine*, *38*(3), 3–4.

Huang, W., Kunfeng, W., Yisheng, L., & FengHua, Z. (2016). Autonomous vehicles testing methods review. *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 163–168). https://doi.org/10.1109/ITSC.2016.7795548

Human Factors and Ergonomics Society. (2020, July 15). *Code of ethics—The Human Factors and Ergonomics Society*. Human Factors and Ergonomics Society. https://www.hfes.org/about-hfes/code-of-ethics

IAEA. (1992). *The Chernobyl accident: Updating of INSAG-1* (Text INSAG-1). IAEA. https://www.iaea.org/publications/3786/the-chernobyl-accident-updating-of-insag-1

Igroup Project Consortium. (2015). Igroup Presence Questionnaire. http://www.igroup.org/pq/ipq/download.php#English

IJsselsteijn, W. A., de Ridder, H., Freeman, J., & Avons, S. E. (2000). Presence: Concept, determinants, and measurement. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 3959*, 520–529. https://doi.org/10/fgk32n

Inagaki, T. (2003). Adaptive automation: Sharing and trading of control. In E. Hollnagel (Ed.), *Handbook of cognitive task analysis* (pp. 147–169). Mahwah, NJ: Lawrence Erlbaum Associates.

Inagaki, T., & Sheridan, T. B. (2019). A critique of the SAE conditional driving automation definition, and analyses of options for improvement. *Cognition, Technology & Work, 21*(4), 569–578.

Jansen, R. J., van der Kint, S. T., & Hermens, F. (2020). Does agreement mean accuracy? Evaluating glance annotation in naturalistic driving data. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01446-9

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Jeon, C. (2015). The virtual flier: The link trainer, flight simulation, and pilot identity. *Technology and Culture, 56*(1), 28–53. https://doi.org/10.1353/tech.2015.0017

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johansson, G., & Rumar, K. (1971). Drivers' brake reaction times. *Human Factors, 13*(1), 23–27. https://doi.org/10/gfkwdm

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science, 4*(1), 71–115. https://doi.org/10/b9j9k9

Jonah, B. A. (1986). Accident risk and risk-taking behaviour among young drivers. *Accident Analysis & Prevention, 18*(4), 255–271. https://doi.org/10/chdx6p

Jones, M. L., & Meurer, K. (2016). Can (and should) Hello Barbie keep a secret? *2016 IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)* (pp. 1–6). https://doi.org/10.1109/ETHICS.2016.7560047

Jung, M. F., Lee, J. J., DePalma, N., Adalgeirsson, S. O., Hinds, P. J., & Breazeal, C. (2013). Engaging robots: Easing complex human-robot teamwork using backchanneling. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 1555–1566). https://doi.org/10.1145/2441776.2441954

Kaber, D. B., Omal, E., & Endsley, M. R. (1999). Level of automation effects on telerobot performance and human operator situation awareness and subjective workload. *Automation Technology and Human Performance: Current Research and Trends,* 165–170.

Kaber, D. B., Riley, J. M., Tan, K.-W., & Endsley, M. R. (2001). On the design of adaptive automation for complex systems. *International Journal of Cognitive Ergonomics, 5*(1), 37–57. https://doi.org/10.1207/S15327566IJCE0501_3

Kaplan, A. D., Kessler, T. T., Sanders, T. L., Cruit, J., Brill, J. C., & Hancock, P. A. (2020). A time to trust: Trust as a function of time in human-robot interaction. In *Trust in Human-Robot Interaction* (pp. 143–157). San Diego, CA: Academic Press.

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford: Oxford University Press.

Kelly, M., Sinha, A., Namkoong, H., Tedrake, R., & Duchi, J. C. (2018). Scalable end-to-end autonomous vehicle testing via rare-event simulation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems, 31* (pp. 9827–9838). Curran Associates, Inc. http://papers.nips.cc/paper/8189-scalable-end-to-end-autonomous-vehicle-testing-via-rare-event-simulation.pdf

Kempton, W. (1986). Two theories of home heat control. *Cognitive Science, 10*(1), 75–90. https://doi.org/10.1207/s15516709cog1001_3

Kempton, W. (1987). Variation in folk models and consequent behavior. *American Behavioral Scientist, 31*(2), 203–218. https://doi.org/10.1177/000276487031002006

Khademi, A., & Honavar, V. (2019). Algorithmic bias in recidivism prediction: A causal perspective. *ArXiv:1911.10640 [Cs, Stat]*. http://arxiv.org/abs/1911.10640

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science, 8*(3), 255–273. https://doi.org/10/cfvbzw

Kirkpatrick, K. (2016). Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Communications of the ACM, 59*(10), 16–17. https://doi.org/10.1145/2983270

Klien, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems, 19*(6), 91–95. https://doi.org/10.1109/MIS.2004.74

Kline, D. W., Kline, T. J. B., Fozard, J. L., Kosnik, W., Schieber, F., & Sekuler, R. (1992). Vision, aging, and driving: The problems of older drivers. *Journal of Gerontology, 47*(1), P27–P34. https://doi.org/10/gfkwcn

Kopec, D., & Tamang, S. (2007). Failures in complex systems: Case studies, causes, and possible remedies. *ACM SIGCSE Bulletin, 39*(2), 180–184. https://doi.org/10.1145/1272848.1272905

Körber, M., Cingel, A., Zimmermann, M., & Bengler, K. (2015). Vigilance decrement and passive fatigue caused by monotony in automated driving. *Procedia Manufacturing, 3*, 2403–2409. https://doi.org/10/gfkwd3

Kozak, J. J., Hancock, P. A., Arthur, E. J., & Chrysler, S. T. (1993). Transfer of training from virtual reality. *Ergonomics, 36*(7), 777–784. https://doi.org/10.1080/00140139308967941

Lacey, C., & Caudwell, C. (2019). Cuteness as a 'Dark Pattern' in home robots. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 374–381) https://doi.org/10.1109/HRI.2019.8673274

Lau, N., Fridman, L., Borghetti, B. J., & Lee, J. D. (2018). Machine learning and human factors: Status, applications, and future directions, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 62*(1), 135–138.

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science, 19*(3), 143–148. https://doi.org/10.1177/0963721410370295

Lebedev, A. (2011, July 21). *The Man Who Saved the World* finally recognized. MosNews. https://web.archive.org/web/20110721000030/http://www.worldcitizens.org/petrov2.html

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(1), 50–80. https://doi.org/10/dr6jf9

Lee, K. M. (2004a). Presence, explicated. *Communication Theory, 14*(1), 27–50. https://doi.org/10/cw7f7f

Lee, K. M. (2004b). Why presence occurs: Evolutionary psychology, media equation, and presence. *Presence: Teleoperators and Virtual Environments, 13*(4), 494–505. https://doi.org/10/bd8dqd

Lerner, N. D. (1993). Brake perception-reaction times of older and younger drivers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 37*(2), 206–210. https://doi.org/10/gfkwcq

Li, J., Cho, M.-J., Xuan, Z., Malle, B. F., & Ju, W. (2016, April 5). From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with Self-driving cars. SAE 2016 World Congress. http://papers.sae.org/2016-01-0164/

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics HFE-1,* (1), 4–11. https://doi.org/10.1109/THFE2.1960.4503259

Lin, P. (2017, April 5). Here's how Tesla solves a self-driving crash dilemma. *Forbes*. https://www.forbes.com/sites/patricklin/2017/04/05/heres-how-tesla-solves-a-self-driving-crash-dilemma/

Lopez, J. (2019, July 11). GM Super Cruise rumored to receive level 3 upgrades. GM Authority. http://gmauthority.com/blog/2019/07/gm-super-cruise-rumored-to-receive-level-3-upgrades/

Lorenz, B., Di Nocera, F., Röttger, S., & Parasuraman, R. (2001). The effects of level of automation on the out-of-the-loop unfamiliarity in a complex dynamic fault-management task during simulated spaceflight operations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 45*(2), 44–48.

Louw, T., Kuo, J., Romano, R., Radhakrishnan, V., Lenné, M. G., & Merat, N. (2019). Engaging in NDRTs affects drivers' responses and glance patterns after silent automation failures. *Transportation Research Part F: Traffic Psychology and Behaviour, 62*, 870–882. https://doi.org/10.1016/j.trf.2019.03.020

Mackworth, J. F. (1969). *Vigilance and habituation: A neuropsychological approach*. New York: Penguin.

Mackworth, J. F. (1970). *Vigilance and attention: A signal detection approach*. New York: Penguin.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology, 1*(1), 6–21. https://doi.org/10/dqn8q5

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124). https://doi.org/10/gfkwb3

Manyika, J. (2017). A future that works: AI, automation, employment, and productivity. McKinsey Global Institute Research, Technical Report, 60.

Maule, A. J., Hockey, G. R. J., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: Changes in affective state and information processing strategy. *Acta Psychologica, 104*(3), 283–301. https://doi.org/10/dm25cf

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review, 20*(3), 709. https://doi.org/10/fs6wzz

McGehee, D. V., Brewer, M., Schwarz, C., & Walker Smith, B. (2016). Review of automated vehicle technology: Policy and implementation implications (MATC-MU:276). University of Iowa. https://rosap.ntl.bts.gov/view/dot/30702

McKnight, A. J., & McKnight, A. S. (2003). Young novice drivers: Careless or clueless? *Accident Analysis & Prevention, 35*(6), 921–925. https://doi.org/10.1016/S0001-4575(02)00100-8

Merkow, M. S., & Breithaupt, J. (2014). *Information security: Principles and practices*. Harlow: Pearson Education.

Meyer, J., & Lee, J. D. (2013). *Trust, reliance, and compliance*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199757183.013.0007

Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995). Augmented reality: A class of displays on the reality-virtuality continuum. *SPIE, 2351,* 282–292.

Miller, D. (2016). AgentSmith: Exploring agentic systems. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 234–238). https://doi.org/10/gfkwdx

Miller, D., & Ju, W. (2015). Joint cognition in automated driving: Combining human and machine intelligence to address novel problems. *2015 AAAI Spring Symposium Series*. http://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10308

Miller, J., Ward, C., Lee, C., D'Ambrosio, L., & Coughlin, J. (2018). Sharing is caring: The potential of the sharing economy to support aging in place. *Gerontology & Geriatrics Education, 0*(0), 1–23. https://doi.org/10.1080/02701960.2018.1428575

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 279–288.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies, 45*(6), 669–678. https://doi.org/10.1006/ijhc.1996.0073

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem." *Emotion, 12*(2), 364–370. https://doi.org/10/fcqpqh

Navon, D., & Gopher, D. (1978). Interpretations of task difficulty in terms of resources: Efficiency, load, demand, and cost composition. (No. ADA070937). Technion-Israel Institute of Technology. https://apps.dtic.mil/dtic/tr/fulltext/u2/a070937.pdf

Nebeker, C., Torous, J., & Ellis, R. J. B. (2019). Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Medicine, 17*(1), 137.

Neisser, U. (1976). General, academic, and artificial intelligence. *The Nature of Intelligence, 135*, 144.

Nicolelis, M. (2011). *Beyond boundaries: The new neuroscience of connecting brains with machines—and how it will change our lives*. Basingstoke: Macmillan.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics, 2*(1), 25–42. https://doi.org/10/dn6p6h

Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review, 75*(6), 522–536. https://doi.org/10.1037/h0026699

Norman, D. A. (1983). Some observations on mental models. *Mental Models, 7*(112), 7–14.

Nyholm, S., & Frank, L. E. (2017). From sex robots to love robots: Is mutual love with a robot possible? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 12.

Nyholm, S., & Smids, J. (2020). Can a robot be a good colleague? *Science and Engineering Ethics, 26*(4), 2169–2188.

Oh, C., Song, J., Choi, J., Kim, S., Lee, S., & Suh, B. (2018). I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* 1–13. https://doi.org/10.1145/3173574.3174223

Openwater. (2020). Openwater. https://www.openwater.cc

Orne, M. T., & Holland, C. H. (1968). On the ecological validity of laboratory deceptions. *International Journal of Psychiatry, 6*, 282–293.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE*

Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10/c6zf92

Peña Gangadharan, S., Eubanks, V., & Barocas, S. (Eds.). (2014). *Data and discrimination: Collected essays*. Open Technology Institute. https://rws511.pbworks.com/w/file/fetch/88176947/OTI-Data-an-Discrimination-FINAL-small.pdf

Pernice, K. (2015, February 22). Emotional design fail: Divorcing my nest thermostat. Nielsen Norman Group. https://www.nngroup.com/articles/emotional-design-fail/

Philip, P., Sagaspe, P., Moore, N., Taillard, J., Charles, A., Guilleminault, C., & Bioulac, B. (2005). Fatigue, sleep restriction and driving performance. *Accident Analysis & Prevention, 37*(3), 473–478. https://doi.org/10.1016/j.aap.2004.07.007

PhishMe. (2015). Enterprise phishing susceptibility report. PhishMe. https://cofense.com/wp-content/uploads/2017/10/PhishMe_EnterprisePhishingSusceptibilityReport_2015_Final.pdf

Posey, C., Roberts, T. L., Lowry, P. B., Bennett, R. J., & Courtney, J. F. (2013). Insiders' Protection of organizational information assets: Development of a systematics-based taxonomy and theory of diversity for protection-motivated behaviors. *MIS Quarterly, 37*(4), 1189–1210.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General, 109*(2), 160–174. https://doi.org/10.1037/0096-3445.109.2.160

Proud, R. W., Hart, J. J., & Mrozinski, R. B. (2003). Methods for determining the level of autonomy to design into a human spaceflight vehicle: A function specific approach. Lyndon B Johnson Center, Houston, TX: National Aeronautics and Space Administration.

Raisch, S., & Krakowski, S. (2020). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 12*.

Ramachandran, V. S. (1998). Consciousness and body image: Lessons from phantom limbs, Capgras syndrome and pain asymbolia. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 353*(1377), 1851–1859. https://doi.org/10.1098/rstb.1998.0337

Ramachandran, V. S. (2012). *The tell-tale brain: A neuroscientist's quest for what makes us human*. New York: W. W. Norton & Company.

Rasmussen, J. (1986). *Information processing and human-machine interaction: An approach to cognitive engineering*. Amsterdam: North-Holland.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge University Press. http://www.humanityonline.com/docs/the%20media%20equation.pdf

Reeves, B., Ram, N., Robinson, T. N., Cummings, J. J., Giles, C. L., Pan, J., Chiatti, A., Cho, M., Roehrick, K., Yang, X., Gagneja, A., Brinberg, M., Muise, D., Lu, Y., Luo, M., Fitzgerald, A., & Yeykelis, L. (2019). Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them. *Human–Computer Interaction, 0*(0), 1–52. https://doi.org/10.1080/07370024.2019.1578652

Reichardt, C. S. (2011). Criticisms of and an alternative to the Shadish, Cook, and Campbell validity typology. *New Directions for Evaluation, 2011*(130), 43–53. https://doi.org/10.1002/ev.364

Richardson, G. P., Andersen, D. F., Maxwell, T. A., & Stewart, T. R. (1994). Foundations of mental model research. *Proceedings of the 1994 International System Dynamics Conference* (pp. 181–192). http://www.albany.edu/~gpr/MentalModels.pdf

Riley, V. (1989). A general model of mixed-initiative human-machine systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 33*(2), 124–128. https://doi.org/10.1177/154193128903300227

Roelfsema, P. R., Denys, D., & Klink, P. C. (2018). Mind reading and writing: the future of neurotechnology. *Trends in Cognitive Sciences, 22*(7), 598–610. https://doi.org/10.1016/j.tics.2018.04.001

Rogers, E. M., & Bhowmik, D. K. (1970). Homophily-heterophily: relational concepts for communication research. *Public Opinion Quarterly, 34*(4), 523. https://doi.org/10.1086/267838

SAE International. (2016). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (No. J3016A). SAE International. http://standards.sae.org/j3016_201609/

Salvucci, D. D. (2013). *Multitasking*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199757183.013.0004

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review, 115*(1), 101–130. https://doi.org/10.1037/0033-295X.115.1.101

Salvucci, D. D., & Taatgen, N. A. (2010). *The multitasking mind*. Oxford: Oxford University Press.

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors, 37*(1), 5–19. https://doi.org/10/dj23nb

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1926–1943). Hoboken, NJ: Wiley.

Sauer, J., Nickel, P., & Wastell, D. (2013). Designing automation for complex work environments under different levels of stress. *Applied Ergonomics, 44*(1), 119–127.

Sawyer, B. D., Dobres, J., Chahine, N., & Reimer, B. (2017). The cost of cool: Typographic style legibility in reading at a glance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1), 833–837.

Sawyer, B. D., Dobres, J., Chahine, N., & Reimer, B. (2020). The great typography bake-off: Comparing legibility at-a-glance. *Ergonomics, 63*(4), 391–398.

Sawyer, B. D., Finomore, V. S., Calvo, A. A., & Hancock, P. A. (2014). Google Glass: A driver distraction cause or cure? *Human Factors, 56*(7), 1307–1321.

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

Sawyer, B. D., Mehler, B., & Reimer, B. (2017). Toward an antiphony framework for dividing tasks into subtasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Sawyer, B. D., Wolfe, B., Dobres, J., Chahine, N., Mehler, B., & Reimer, B. (2020). Glanceable, legible typography over complex backgrounds. *Ergonomics,* 1–20.

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In M. Mustapha & P. A. Hancock (Eds.), *Human performance in automated and autonomous systems: Current theory and methods* (pp. 103–149). Boca Raton, FL: CRC Press.

Schaefer, K. E., Adams, J. K., Cook, J. G., Bardwell-Owens, A., & Hancock, P. A. (2015). The future of robotic design: Trends from the history of media representations. *Ergonomics in Design, 23*(1), 13–19.

Sebescen, N., & Vitak, J. (2017). Securing the human: Employee security vulnerability risk in organizational settings. *Journal of the Association for Information Science and Technology, 68*(9), 2237–2247. https://doi.org/10.1002/asi.23851

Shahriari, K., & Shahriari, M. (2017). IEEE standard review — Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)* (pp. 197–201). https://doi.org/10.1109/IHTC.2017.8058187

Shekhar, S. S. (2019). Artificial intelligence in automation. *Artificial Intelligence, 3085*(06), 14–17.

Sheridan, T. B. (1975). Considerations in modeling the human supervisory controller. *IFAC Proceedings Volumes, 8*(1, Part 3), 223–228. https://doi.org/10/gfkwdv

Sheridan, T. B. (1992a). *Telerobotics, automation, and human supervisory control*. Cambridge, MA: MIT Press.

Sheridan, T. B. (1992b). Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments,* 1(1), 120–126. https://doi.org/10/gdcftg

Sheridan, T. B. (2002). *Humans and automation: System design and research issues*. Human Factors and Ergonomics Society.

Sheridan, T. B. (2006). Supervisory control. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (3rd ed., pp. 1025–1052). Hoboken, NJ: Wiley. https://doi.org/10.1002/0470048204.ch38

Sheridan, T. B., & Hennessy, R. T. (1984). *Research and modeling of supervisory control behavior: Report of a workshop*. Washington, DC: National Research Council Committee on Human Factors. https://apps.dtic.mil/sti/citations/ADA149621

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics, 1*(1), 89–129. https://doi.org/10.1518/155723405783703082

Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators. NASA Technical Reports Server.

Shneiderman, B. (1998). Codex, Memex, Genex: The pursuit of transformational technologies. *International Journal of Human–Computer Interaction, 10*(2), 87–106. https://doi.org/10.1207/s15327590ijhc1002_1

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal, 31*(2), 47–53.

Simon, H. A. (1965). *The shape of automation for men and management* (1st ed.). New York: Harper & Row Publishers, Inc.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059–1074. https://doi.org/10.1068/p281059

Sirkin, D., Martelaro, N., Johns, M., & Ju, W. (2017). Toward measurement of situation awareness in autonomous vehicles. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 405–415). https://doi.org/10.1145/3025453.3025822

Skrypchuk, L., Langdon, P., Sawyer, B. D., & Clarkson, P. J. (2020). Unconstrained design: Improving multitasking with in-vehicle information systems through enhanced situation awareness. *Theoretical Issues in Ergonomics Science, 21*(2), 183–219.

Skrypchuk, L., Langdon, P., Sawyer, B. D., Mouzakitis, A., & Clarkson, P. J. (2019). Enabling multitasking by designing for situation awareness within the vehicle environment. *Theoretical Issues in Ergonomics Science, 20*(2), 105–128.

Slater, M., & Usoh, M. (1993). Presence in immersive virtual environments. In *Proceedings of IEEE Virtual Reality Annual International Symposium* (pp. 90–96). https://doi.org/10.1109/VRAIS.1993.380793

Smith, A., & Anderson, M. (2018). Social media use in 2018. *Pew Research Center, 1*, 1–4.

Smith, K. (1989). Computer security-threats, vulnerabilities, and countermeasures. *Information Age, 11*(4), 205–210.

Smith, K., & Hancock, P. A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors, 37*(1), 137–148.

Sommer, D., & Golz, M. (2010). Evaluation of PERCLOS based current fatigue monitoring technologies. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* (pp. 4456–4459). https://doi.org/10.1109/IEMBS.2010.5625960

Stark, L. (2016). The emotional context of information privacy. *The Information Society, 32*(1), 14–27. https://doi.org/10.1080/01972243.2015.1107167

Sujan, M., White, S., Furniss, D., Habli, I., Grundy, K., Grundy, H., Nelson, D., Elliott, M., & Reynolds, N. (2019). Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health and Care Informatics*.

Sütfeld, L. R., Gast, R., König, P., & Pipa, G. (2017). Using virtual reality to assess ethical decisions in road traffic scenarios: Applicability of value-of-life-based models and influences of time pressure. *Frontiers in Behavioral Neuroscience, 11*. https://doi.org/10/gfkwc2

Svenson, O., & Maule, A. J. (Eds.). (1993). *Time pressure and stress in human judgment and decision making*. New York: Plenum Press.

Takayama, L., Groom, V., & Nass, C. (2009). I'm sorry, Dave: I'm afraid I won't do that: Social aspects of human-agent conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099–2108). https://doi.org/10/dpkr3w

Takayama, L., Ju, W., & Nass, C. (2008). Beyond dirty, dangerous and dull: What everyday people think robots should do. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 25–32). https://doi.org/10.1145/1349822.1349827

Taylor, M. (2016, October 7). Self-driving Mercedes-Benzes will prioritize occupant safety over pedestrians. *Car and Driver*. http://blog.caranddriver.com/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness* (Rev. and expanded ed.). New York: Penguin.

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. *ArXiv Preprint ArXiv:2007.05558*.

Uber. (2019, December). Uber's US safety report. https://www.uber.com/us/en/about/reports/us-safety-report/

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology, 20*(1), 27–40.

van der Vecht, B., van Diggelen, J., Peeters, M., Barnhoorn, J., & van der Waa, J. (2018). SAIL: A social artificial intelligence layer for human-machine teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 262–274.

van Maanen, P.-P., Lindenberg, J., & Neerincx, M. A. (2005). Integrating human factors and artificial intelligence in the development of human-machine cooperation. *Proceedings of the 2005 International Conference on Artificial Intelligence*. 2005 International Conference on Artificial Intelligence (ICAI'05).

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2012). Trust in smart systems sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 54*(5), 799–810. https://doi.org/10.1177/0018720812443825

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2015). Trusting a virtual driver that looks, acts, and thinks like you. *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 0018720815580749. https://doi.org/10/gfkv99

Vieane, A., Funke, G., Gutzwiller, R., Mancuso, V., Sawyer, B. D., & Wickens, C. (2016). Addressing human factors gaps in cyber defense. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 770–773.

Vincent, J. (2018, January 12). Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. *The Verge*. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

Wallace, S., Treitman, R., Kumawat, N., Arpin, K., Huang, J., Sawyer, B. D., & Bylinskii, Z. (2020a). Towards readability individuation: The right changes to text format make large impacts on reading speed. *Journal of Vision, 20*(10), 17–17.

Wallace, S., Treitman, R., Kumawat, N., Arpin, K., Huang, J., Sawyer, B., & Bylinskii, Z. (2020b). Individual differences in font preference & effectiveness as applied to interlude reading in the digital age. *Journal of Vision*, *20*(11), 412–412.

Waller, P. F. (1991). The older driver. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 33*(5), 499–505. https://doi.org/10/gfkwcm

Waltzman, R. (2017). *The weaponization of information: The need for cognitive security*. Santa Monica, CA: RAND Corporation. https://doi.org/10.7249/CT473

Wang, W., & Siau, K. (2019). Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity:

A review and research agenda. *Journal of Database Management (JDM), 30*(1), 61–79.

Warwick, K. (2016). Transhumanism: Some practical possibilities. *FIfF-Kommunikation. Zeitschrift für Informatik und Gesellschaft, 2*, 24–25.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2020). "Let me explain!": Exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces,* 1–12.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors, 50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wickens, C. D. (2013). *Attention*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199757183.013.0003

Wickens, C. D., & Gutzwiller, R. S. (2017). The status of the strategic task overload model (STOM) for predicting multi-task management. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1), 757–761. https://doi.org/10.1177/1541931213601674

Wickens, C. D., Gutzwiller, R. S., Vieane, A., Clegg, B. A., Sebok, A., & Janes, J. (2016). Time sharing between robotics and process control: Validating a model of attention switching. *Human Factors, 58*(2), 322–343. https://doi.org/10.1177/0018720815622761

Wickens, C. D., Santamaria, A., & Sebok, A. (2013). A computational model of task overload management and task switching. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 57*(1), 763–767. https://doi.org/10.1177/1541931213571167

Wiener, N., & Heims, S. J. (1989). *The human use of human beings: Cybernetics and society*. New York: Free Association Books.

Wilson, J. R., & Rutherford, A. (1989). Mental models: Theory and application in human factors. *Human Factors, 31*(6), 617–634. https://doi.org/10.1177/001872088903100601

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems. *Proceedings of the IEEE, 107*(3), 509–517.

Wintersberger, P., Frison, A.-K., Riener, A., & Hasirlioglu, S. (2017). The experience of ethics: Evaluation of self harm risks in automated vehicles. In *IEEE Intelligence Vehicles Symposium* (vol. 28, pp. 385–391). Redondo Beach, CA.

Wolfe, B., Sawyer, B. D., & Rosenholtz, R. (2020). Toward a theory of visual information acquisition in driving. *Human Factors*, 0018720820939693.

Woods, D. (1994). Automation: Apparent simplicity, real complexity. *Human Performance in Automated Systems:* Current Research and Trends*,* 1–7.

Yeykelis, L., Cummings, J. J., & Reeves, B. (2014). Multitasking on a single device: Arousal and the frequency, anticipation, and prediction of switching between media content on a computer: multitasking and arousal. *Journal of Communication, 64*(1), 167–192. https://doi.org/10/f57c3z

Yeykelis, L., Cummings, J. J., & Reeves, B. (2017). The fragmentation of work, entertainment, e-mail, and news on a personal computer: motivational predictors of switching between media content. *Media Psychology, 0*(0), 1–26. https://doi.org/10/gfkwdj

Yoo, C. W., Sanders, G. L., & Cerveny, R. P. (2018). Exploring the influence of flow and psychological ownership on security education, training and awareness effectiveness and security compliance. *Decision Support Systems, 108*, 107–118. https://doi.org/10.1016/j.dss.2018.02.009

Young, M. S., & Stanton, N. A. (2002a). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science, 3*(2), 178–194. https://doi.org/10.1080/14639220210123789

Young, M. S., & Stanton, N. A. (2002b). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 44*(3), 365–375. https://doi.org/10/d6m77b

Young, M. S., & Stanton, N. A. (2006). The decay of malleable attentional resources theory. In P. D. Bust (Ed.), *Contemporary ergonomics* (pp. 253–257). London: Taylor & Francis. https://www.researchgate.net/profile/Neville_Stanton/publication/289423242_The_decay_of_malleable_attentional_resources_theory/links/569ffb3a08ae4af52546db31.pdf

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. *ArXiv Preprint ArXiv:1812.02953*.

Zimmermann, A., & Di Rosa, E. (2019, December 12). Technology can't fix algorithmic injustice. *Boston Review*. http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic.