FRAMING HUMAN-AUTOMATION REGULATION: A NEW MODUS OPERANDI FROM COGNITIVE ENGINEERING

Marc C. Canellas[*]*, Matthew J. Miller, Yosef S. Razin, Dev Minotra, Raunak Bhattacharyya, and Rachel A. Haga*

Human-automated systems are becoming ubiquitous in our society, from the one-on-one interactions of a driver and their automated vehicle to large-scale interactions of managing a world-wide network of commercial aircraft. Realizing the importance of effectively governing these human-automated systems, there been a recent renaissance of legal-ethical analysis of robotics and artificial-intelligence-based systems. As cognitive engineers, we authored this paper to embrace our responsibility to support effective governance of these human-automated systems. We believe that there are unique synergies between the cognitive engineers who shape human-automated systems by designing the technology, training, and operations, and the lawyers who design the rules, laws, and governance structures of these systems. To show how cognitive engineering can provide a foundation for effective governance, we define and address five essential questions regarding human-automated systems: 1) Complexity: What makes human-automation systems complex? 2) Definitions: How should we define and classify different types of human-autonomous systems? 3) Transparency: How do we determine and achieve the right levels of transparency for operators and regulators? 4) Accountability: How should we determine responsibility for the actions of human-automation systems? 5) Safety: How do human-automated systems fail? Our answers, drawn from the diverse domains related to cognitive engineering, show that care should be taken when making assumptions about human-automated systems, that cognitive engineering can provide a strong foundation for legal-ethical regulations of human-automated systems, and that there is still much work to be done by lawyers, ethicists, and technologists together.

---

[*]All authors are from the Cognitive Engineering Center (CEC) at the Georgia Institute of Technology, Atlanta, GA, USA (http://www.cec.gatech.edu). M. Canellas, R. Haga, M. Miller, and R. Bhattacharyya are Ph.D. candidates in aerospace engineering at the School of Aerospace Engineering. Y. Razin is a Ph.D. candidate in robotics at the Institute for Robotics and Intelligent Machines. Dr. D. Minotra, is a postdoctoral fellow at the School of Aerospace Engineering. After M. Canellas, authors are ordered chronologically by their primary section: M. Miller (Complexity), M. Canellas (Definitions), Y. Razin (Transparency), D. Minotra and R. Bhattacharyya (Accountability), and R. Haga (Safety). This paper was submitted and presented at WeRobot 2017 in New Haven, CT, on April 1, 2017.

TABLE OF CONTENTS

I. INTRODUCTION

Human-automation systems are everywhere. They are the driver monitoring the highly-automated vehicle as it navigates through a traffic light. They are the teams of humans and algorithms determining whether a suspect's photo matches an FBI database. They are the squadron of commanders, pilots, and soldiers operating drones over a military warzone. They are the thousands of pilot-autopilot teams interacting with each other and thousands of air traffic controllers to ensure millions of people fly safely and efficiently each day. Even you, the reader, are part of a human-automation system.

Legal scholars, ethical scholars, and policy makers have recently embraced their role in governing these human-automation systems, and rightfully so. Technology no longer seems to have a limit on its capabilities. The question is no longer what can be done with technology like artificial intelligence and robotics, but what should be done with them (Marchant, Abbott, & Allenby, 2014; Marchant, Allenby, & Herkert, 2011). Therefore, though legal-ethical perspectives on human-automation systems have been around as long as the technologies themselves (Calo, 2016), there has suddenly been a renaissance of legal-ethical analysis in recent years: IEEE's Global Initiative for Ethically Aligned Design (2016a), IBM's Principles for the Cognitive Era (2017), the tenants of the Partnership on AI (Partnership on AI, 2017) , Future of Life Institute's Asilomar AI Principles (Future of Life Institute, 2017), and the One Hundred Year Study of Artificial Intelligence (Grosz et al., 2016). Formal lawmaking has finally started to guide the next generation of these systems whether they be artificial intelligence (Holdren & Smith, 2016), autonomous vehicles (NHTSA, 2016), or autonomous weapons (UN CCW, 2016).

As cognitive engineers, we authored this paper to embrace our responsibility to support effective governance of these human-automated systems. Cognitive engineering specializes in the design, analysis, and evaluation of complex, sociotechnical, and safety-critical systems that are dependent on human-automation interaction. As its practitioners, we are uniquely capable in providing the necessary grounding for the construction of practical legal and ethical frameworks.

A review of many legal and ethical frameworks reveals that many of their concerns and objectives can be evaluated through cognitive engineering's experience in designing, analyzing, and evaluating human-automation systems. For instance, some assumptions of the legal and ethical frameworks

are flawed: that a human-in-the-loop is inherently a safety solution; that more transparency will engender more trust; that increased automation results in increased safety; or that accountability can be guaranteed. Conversely, many of their goals can be achieved with the support of cognitive engineering: implementing the norms and values of communities within human-automation systems; ensuring more robust, reliable, and trustworthy human-automation systems; and developing certification and accountability frameworks for human-automation systems.

The key synergy between the cognitive engineering and legal communities is that both try to shape societal behavior – only their methods and perspectives differ. Cognitive engineering shapes social and technical systems (referred to as sociotechnical systems) by designing technology, training, and concepts of operations, whereas the legal community designs the rules, laws, and governance structures. Cognitive engineering focuses on understanding the formal and informal work and the constraints existing within human-automated systems whereas the legal community has methods of constraining and defining work. To us, the cognitive engineering efforts that promote enhanced human-automated system performance are analogous to many of the efforts found in the legal community (Canellas & Haga, 2015, 2016).

In this paper, we define and address five main questions that we believe are general enough to be the starting point for governance of any type of human-automation system:

- Complexity: What makes human-automation systems complex?
- Definitions: How should we define and classify different types of human-autonomous systems?
- Transparency: How do we determine and achieve the right levels of transparency for operators and regulators?
- Accountability: How should we determine responsibility for the actions of human-automation systems?
- Safety: How do human-automated systems fail?

These provide a pragmatic foundation for constructing new legal-ethical frameworks that will govern these systems. The five questions are addressed in this paper from a critical literature review of fields within the broad scope of cognitive engineering, including human factors and ergonomics, system safety, human cognition and behavior, robotics, and computer science.

Throughout, our intention in addressing these questions is practical: we wish to provide a foundation of key concerns, models, theories, and frameworks that can help those governing these systems to ask the right questions within their own specific applications. Each of the five questions build upon each other to characterize human-automation systems. First, we grapple with the fundamental question of complexity: what are human-autonomous systems and what makes them so difficult to understand? Given this complexity, how should we appropriately define and classify such systems (definitions) or ensure that the humans interacting with the system understand what the automation is doing (transparency). Finally, when the human-automation interaction inevitably breaks down it may result in damages, accidents, and even injuries, raising the questions of how the failure occurred (safety) and who or what is responsible (accountability).

Instead of taking a stance that automation is inherently desirable or nefarious, our answers to these questions provide the foundation for appropriate skepticism and optimism with respect to human-automation systems. There are particular myths both against and in favor of automation that must be dispelled (Bradshaw, Hoffman, Johnson, & Woods, 2013). But there are also virtues of implementing automation that can make society operate in a safer and more effective manner (Johnson, Bradshaw, Hoffman, Feltovich, & Woods, 2014). We contend that cognitive engineering offers one path toward achieving this balance.

We do not pretend that this paper can answer all the questions and concerns of lawyers and ethicists related to human-automation systems. In fact, as designers and evaluators of these systems, we don't have all the answers ourselves. The only universal truth from the literature is that there is no such thing as a perfect human-automation system. There is no perfect human-automation interaction with perfectly safe performance, perfectly transparent communication, and perfect delineation of human and automation roles and responsibilities. For each operating context, the governance structure will be unique: the classifications, the rules of transparency, the acceptable levels of safety, and the accountability frameworks. Lawyers and ethicists, together with technologists, must work together under a common framework(s) to ensure that each iteration of each human-automation system is safer, more effective, and better governed than the previous.

II. COMPLEXITY

> *"Our fascination with new technologies is based on the assumption that more powerful automation will overcome human limitations and make our systems `faster, better, cheaper' resulting in simple, easy tasks for people. But how do new technology and more powerful automation change our work? What [cognitive engineering has] found is not stories of simplification through more automation but stories of complexity and adaptation. … Ironically, more autonomous machines have created the requirement for more sophisticated forms of coordination across people, and across people and machines, to adapt to new demands and pressures."*
>
> – David Woods and Erik Hollnagel (2006a, p. back cover)

We live in a world with ever-growing technological and automated capability. The dynamics and behavior of human-automated systems are becoming less transparent and the legal community's ability to cope with the legal implications of accountability and safety of both intended and unintended consequences is ever more challenging. To address these challenges of definitions, transparency, safety, and accountability, one must start by asking: what makes human-automation systems different from other systems? In other words, what makes them complex? And then, how can all the factors of these complex systems be framed and understood in a meaningful way?

Effective governance of humans and the complex technologies they utilize will require a deep understanding of the interaction between people and the organizational structures they operate within (the social system) and the technologies (the technical systems) they utilize to successfully achieve overall system goals and objectives. These systems involve context-rich workplace settings, organizational structure, human operators, and sophisticated technology that when taken collectively are known as complex sociotechnical systems (Baxter & Sommerville, 2011; Walker, Stanton, Salmon, & Jenkins, 2008; Waterson et al., 2015). This section briefly explores the attributes of human-automation system complexity within the context of sociotechnical systems and reveals some potentially useful methods and insight from the cognitive systems engineering literature (hereafter included in the general term, cognitive engineering), see (Hollnagel & Woods, 1983; Rasmussen, Pejtersen, & Goodstein, 1994; Woods & Hollnagel, 2006b; Woods & Roth, 1988) for a review. In short, cognitive engineering aims to collectively understand and advance the

intersections of people, their work, and the technologies that support and enable both.

Throughout this section, we will emphasize the shared goal of cognitive engineering and the legal community: to appropriately shape the behavior of these sociotechnical systems. Legal frameworks provide specific constraints and requirements on these systems whereas cognitive engineering seeks to understand the constraints and requirements to develop solutions – be they training, technology, or organizational structure. Given cognitive engineering's decades-long focus on human-automated sociotechnical systems, the field has developed ways to understand how complexity manifests itself within a sociotechnical system and formal methods of studying the constraints that shape system behavior (e.g. cognitive work analysis, CWA). Thinking along these lines can be very useful for reasoning through the potential impacts and development of new rules for these human-automation systems. Furthermore, a survey of attempts to implement new technologies in complex sociotechnical systems provides motivation for the necessary skepticism for effective governance (Woods & Hollnagel, 2006a).

A. What makes a human-automation system complex?

1. Nine factors of complexity

Two complimentary cognitive engineering perspectives, shown in Figure 1, offer one way to explore the factors of complexity that are inherent to sociotechnical systems. Examples of these systems include air traffic control, military command and control, and health care, where the constituent elements of people, technologies, and the work they perform collectively contribute to overall system performance. Obviously, these systems can be challenging to quantify and describe, given their multi-faceted nature of operations. To begin understanding how sociotechnical systems operate requires a close examination of the factors that contribute to their complexity. Each of the nine factors, as shown in Figure 1, collectively describe the variety of sources of complexity that impose challenges facing the CSE community.

What makes sociotechnical systems complex is the simple fact that all nine factor levels are simultaneously at play when considering the performance of complex systems. Therefore, the appropriate system boundary must be established by considering how all factors present themselves in the system of interest, in order to understand how they collectively influence the behaviors of the system.

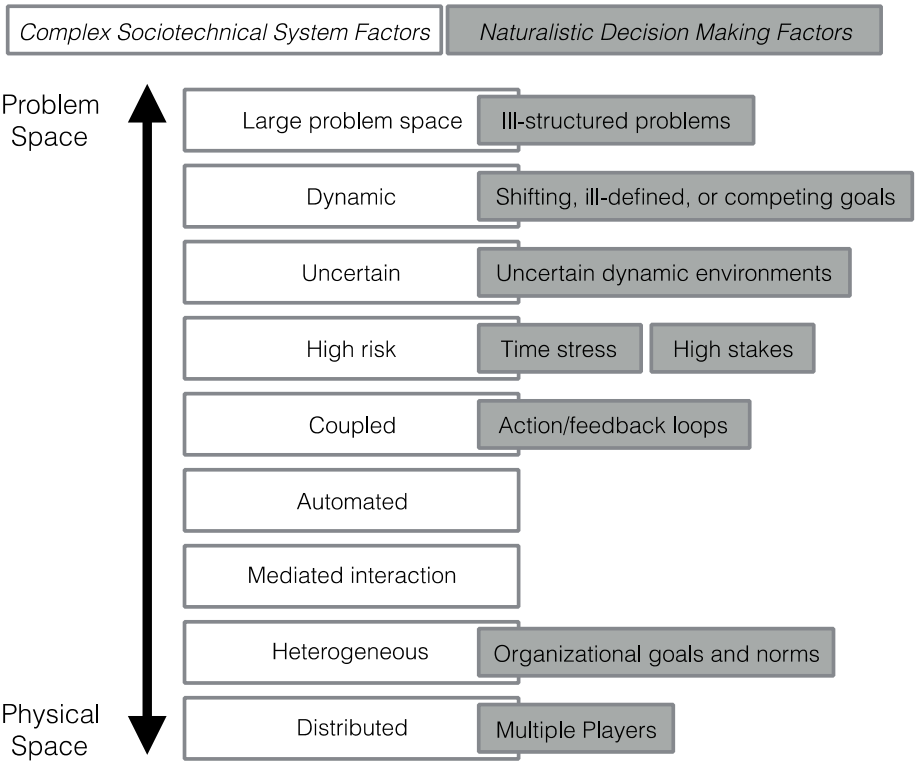

Figure 1. Attributes of complexity considered by the cognitive systems engineering (Vicente, 1999) related to naturalistic decision making factors (Orasanu & Connolly, 1993).

The most obvious factor contributing to system complexity is coping with the sheer volume of problems needing to be solved within human-automation systems. Consider for a moment the range of problems that exist within the U.S. National Airspace System (Felder & Collopy, 2012). When automated aircraft controls were developed in the 1930s, they consisted entirely of electro-mechanical components which could be certified one-by-one with oscilloscopes. Now automation is a complex cyber-physical system incorporating a wide array of functions that is capable of flying a prescribed flight plan from takeoff to landing (Pritchett, 2009). Future visions of automation now entertain concepts of a completely autonomous National Airspace System. The scale of problems illustrated by these various degrees are staggering and remain a fundamental research question (K. Lee, Feron, & Pritchett, 2009). For example, to fully test new software responsible only for automatically delivering flight clearances to an aircraft parked at a gate, "will be literally impossible" using current techniques (Felder & Collopy, 2012, p.

325). What began with automation addressing physical problems (e.g. maintaining altitude), have now become strategic and conceptual (e.g. how should a fleet of aircraft be distributed across the nation to meet customer demands while accounting for weather). Simply put, the more problems that novel automated systems attempt to solve the larger the problem space will become, and the more difficult it will be to capture how any single solution will affect others, or how to integrate the solutions together.

Furthermore, the types of problems that exist can vary from the well-defined (e.g. screw in a bolt or transfer this door from one location to the next) to ill-defined (e.g. make sure this car will preserve passenger safety while driving). Oftentimes the ill-defined portions of problems that exist are left to human operators to cope with (e.g. when an autopilot is outside of its designed mode of operations and the human is required to take over). The presence of difficult and ill-defined problems is compounded by the fact that there typically exists a dynamic and shifting state of operations (Bainbridge, 1997). Problems that a system must address do not remain static and can create conflicts between potential solutions. Adding to the chaos are the dimensions of uncertainty, risk, and the coupled nature of human-automated systems. Many times, these factors are ill-defined and only known or even considered once a system failure has occurred, as described in Sec. VI. Safety.

Two additional factors that contribute to system complexity are the automation itself and the mediated interaction that occurs in human-automated systems. Automation here refers to the capabilities of the automated portion of the system and its specific design characteristics. Even if automation systems could be built to exact design specifications, that does not ensure they are used as intended or disallow the distributed set of operators (human or automated agents) from being negatively impacted by the technology. Note that the automation is only one factor among nine that influence overall complexity. It evidences our concern that regulations focusing specifically on the automation and its capabilities will be unable to address the rest of the many contributing factors to the complexity of the system and the resulting issues of definitions, transparency, safety, and accountability. Automation is not thought about in isolation within the sociotechnical perspective, but rather as part of a collective set of factors. Mediated interaction, through interface design, plays an integral role in how humans perceive and interact with automation. Without a means to effectively perceive and interact, the desired human-automation system performance can be jeopardized.

Finally, factors that contribute to system complexity must involve the consideration of the heterogeneous organization norms and goals of the work-place setting and the actual distribution of actors within the system. Therefore, the cognitive engineering community now favors viewing automated agents as team members within sociotechnical systems who collectively and collaboratively work with humans to achieve goals within the organizational norms and work place setting demands (Johnson, Bradshaw, Feltovich, et al., 2014; Pritchett, 2009)

The collective set of factors in Figure 1 is intended to convey the point that fully comprehending the dimensions of complexity is a difficult, if not impossible, task. But it is important to know what the dimensions are and all the factors affecting the performance of human-automation systems. Because if the dimensions of complexity are ignored, then important considerations immediately become omitted from governance decisions that could have profound impacts on system behavior. The practical advice for lawyers and technologists is to internalize the ideas of each dimension and when faced with a human-automation system try to determine how these nine factors apply.

2. Complexity in future systems

In addition to the factors that contribute to system complexity of existing systems, the notion of complexity becomes even more challenging when considering future systems. Addressing what is known as the envisioned world problem, that is designing future systems to perform future work, brings in four additional perspectives to consider early in the design process as stated below taken from (Woods & Dekker, 2000):

- Plurality: there are multiple versions of how the proposed changes will affect the character of the field of practice in the future.
- Underspecification: each envisioned concept is vague on many aspects of what it would mean to function in that field of practice in the future; in other words, each is a simplification, or partial representation of what it will mean to practice when that envisioned world becomes concrete.
- Ungrounded: envisioned concepts can easily be disconnected or even contradict, from the research base, the actual consequences of the changes on people, technology and work; and

- Overconfident: advocates are miscalibrated and overconfident that, if the systems envisioned can be realized, the predicted consequences and only the predicted consequence will occur.

These same perspectives on designing for the future could be readily applied to new regulations or legislation, which ultimately aim to influence the future operations of complex sociotechnical systems. The application of new constraints must be considered in a future-thinking context where the aforementioned factors can limit the overall impact of what might otherwise be clear intent. Guarding against these known limitations of envisioning the future will need to be addressed by the legal community as they strive towards constraining future systems. As a first step in identifying and striving to understand the contributing components of complexity, lawmakers can more readily apply regulatory intent to shape overall system behavior. In the next section, we propose some useful ways to begin to tackle and model complex systems that account for these factors to make meaningful progress towards system design.

B. How should we frame complexity?

The emphasis on understanding complexity has spurred many advancements in the theory and application of the design of sociotechnical systems, see (Walker et al., 2008) for a review. We highlight here two valuable contributions to this effort for their applicability to the legal community. Born out of the study of nuclear power plant operations, Rasmussen, and later Vicente, developed a framework known as Cognitive Work Analysis (CWA) as a way to capture and articulate the constraints that shape the behaviors of a complex sociotechnical system (Rasmussen et al., 1994; Vicente, 1999) (See (Bisantz & Burns, 2009; Jenkins, Stanton, Salmon, & Walker, 2009) for a more recent examples of CWA applications). The key assumption of CWA is that by making the constraints apparent to agents within the domain, more appropriate strategies and solutions could be made when faced with complex situations. This 'constraint-based' or formative framework emphases an in-depth study of the various constraints of a system and, chiefly, "a demonstration of the various dimensions of the problem" (Rasmussen, cited in Vicente, 1999, p. xi).

Since its inception, CWA has been successfully applied to understand a multitude of complex work domains such as military command and control, nuclear power plant operations, air traffic control, rail transport, and health care. See (Jiancaro, Jamieson, & Mihailidis, 2013; McIlroy & Stanton, 2015; Read, Salmon, & Lenne, 2015) for reviews. The culminating results of these

efforts is a clearer understanding of the work and constraints that shape those domain specific operations. However, some outstanding challenges still face the cognitive engineering community, including providing formal frameworks for guiding the design and development of future systems rather than already-realized systems (M. J. Miller & Feigh, 2017; Read et al., 2015). All too often, technology capabilities drive system development instead of a more concerted effort to understand and design around the desired work functions and objectives of future systems.

Additionally, the field of naturalistic decision making (NDM) has made significant progress to describe how human agents make decisions in real world settings, see (Klein, 2008) for review. Instead of leveraging classical decision making theory, NDM was developed by extensive field studies of experts situated within their work context while making difficult decisions. In doing so, a more detailed depiction was captured of what characteristics of actual work settings and how those characteristics contribute to the decision making capabilities of agents (Orasanu & Connolly, 1993). Moreover, these CWA and NDM investigations already incorporate analyses of the cultural work place and regulatory environment that influence their respective systems operations. In effect, CWA and NDM provide empirically derived approaches to contending with the notions of complexity, previously described in Figure 1.

C. Regulating in a complex, human-automated world

We contend that the factors shown in Figure 1 are an appropriate starting point for contemplating the regulatory environment of complex systems that depend on complex human-automation interaction. Note that automation is only one factor among a multitude of others that are known to shape system behavior. We have ordered and paired the elements from the respective theories to make two additional aspects of these features apparent. First, the factors span the realm of both problem and physical spaces. The physical space refers to the arrangement, distribution, and physical context of the personnel and assets (including advanced technology) within the work domain. The problem space contends with the cognitive demands the work domain demands of its operators. Second, note here that technology is not an explicit factor in Figure 1. Instead, cognitive engineering views technology as a hypothesis about how to best address the problems present within the physical structure of the work domain. Therefore, for effective governance of new complex systems, technology and the automation provided by that technology must be not be evaluated based on human-automation interaction

but rather scrutinized for its ability to promote or hinder the intended work it was built to support. Cognitive engineering promotes such a perspective and provides some methodology that the legal community may find insightful. To provide some context, the following section provides some examples of where cognitive engineering has been applied and met with success.

The ability to make meaningful progress towards synchronizing the demands of complex systems with technological design starts with understanding what types of problems exist as the work is performed. The search, discovery, and formalization of problems within the situated context of those decision-making challenges requires researchers to contend with the variety of contextual features that make the problem space large, dynamic, uncertain, and high risk. Fortunately, examples exist from the CSE community such as large-scale system development efforts that describe how principled mappings between system functional decompositions can be developed in military command and control (Bisantz et al., 2003), identifying traceable links between the results of cognitive analyses and actionable design requirements in the health care informatics development (Hettinger, Roth, & Bisantz, 2017; Jiancaro et al., 2013), synchronizing cockpit display logic with pilot cognitive demands (Riley, DeMers, Misiak, & Schmalz, 1999; Riley, DeMers, Misiak, & Shackleton, 2002; Riley, 1996, 2000), and making military battlefield constraints transparent to commanders (Bennett, Posey, & Shattuck, 2008; Hall, Shattuck, & Bennett, 2012). The consistent thread among these examples is that to yield effective design solutions, the problems must be understood in the context of when decisions must be made and the associated work involved. When this perspective is missed, technology may not be accepted as in the case of digital flight strips being rejected by the French air traffic control community (Mackay, 1999), or worse yet when a highly automated multi-billion dollar air traffic control system is scraped (Britcher, 1999).

We contend that these aforementioned cognitive engineering perspectives of effective system design, which promote a more tenable approach to building desired system performance, in effect are analogous to many of the efforts found in the legal community. To further highlight this linkage, Figure *2* shows the inherent regulatory considerations that exists as part of the CWA framework. The conventional spectrum of system boundaries can extend along the dimensions of the intentional (e.g. the operators' intentions, rules and practices) to the causal (e.g. laws of nature). This spectrum is important because the legal community falls strongly on the intentional end of the spectrum with respect to regulation. In effect, regulation itself is a system boundary that can have profound impacts on the operations and work

performed. The only difference here is that the definition of complex system has expanded to a larger scale. Instead of a complex system being the physical nuclear power plant, the system is the nuclear power plant industry as a whole. What makes this constraint-based perspective challenging is that regulation must effectively be transformed and mapped to meet the eventual specific causal constraints that shape system behavior. Building effective regulation for human-automation systems will require a purposeful endeavor into examining the factors of system complexity and linking intentional and causal constraints to obtain desired system performance. We contend throughout the remainder of this paper that addressing the topics of definitions, transparency, safety, and accountability offer a tenable path toward integrating these cognitive engineering perspectives with the law.

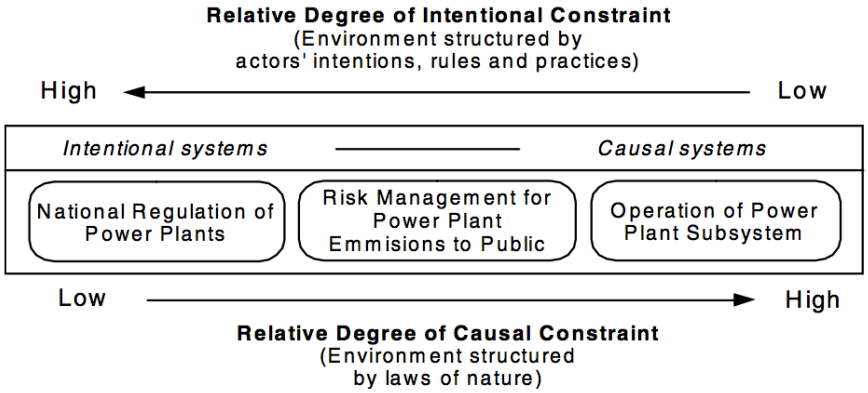

Figure 2. Relative degrees of constraint definition ranging from intentional to causal (Hajdukiewicz, Burns, Vicente, & Eggleston, 1999; Rasmussen et al., 1994).

III. DEFINITIONS

> *"When words are used sloppily, concepts become fuzzy, thinking is muddled, communication is ambiguous, and decisions and actions are suboptimal, to say the least."*
>
> – Stan Stan Kaplan (1990)

Definitions are the foundation for any attempt to regulate or govern. They are essential for distinguishing between different types of systems that will be governed differently because of their differential impact on individuals, organizations, or societies. Informally, answering the question, "What are we talking about?" is often the first point of discussion regarding a new technology or new use. However, with all their sociotechnical complexities as described in Sec. II. Complexity, defining human-automation systems has been and will continue to be a considerably difficult task. Particularly, because our perspectives on automated systems – especially robots – are significantly shaped by the metaphors, stories, and names[1] that we use to describe them (Calo, 2014; Darling, Nandy, & Breazeal, 2015; Darling, 2017; Richards & Smart, 2013).

As the law attempts to define and classify these human-automation systems, debate and confusion inevitably reign, often stymying substantive discussion about transparency, safety, and accountability. For example, the three consecutive meetings of experts on autonomous weapons systems at the United Nations from 2014 to 2016 made little progress on meaningful human control, accountability, and weapons reviews (Ford & Jenks, 2016; Knuckey, 2014), largely because of the lack of consensus on defining "autonomy" and "autonomous weapons systems", as stated by the 2016 U.S. delegation's statement (Meier, 2016).

In other situations, the lack of clear definitions and classifications can leave linguistic loopholes within the rules for others to exploit. In 2016, Uber deployed self-driving cars in San Francisco despite the California Department of Motor Vehicles requiring the cars to be registered and regulated as "autonomous vehicles." Uber claimed their cars did not meet the

[1]As compiled in Canellas & Haga (2015, p. 1), autonomous weapons systems have many names colored by the authors' perspective on the technology: autonomous weapons systems (AWS, U.S. Department of Defense, 2012), lethal autonomous weapons systems (Hagerott, 2014), lethal autonomous robots (Arkin, 2013; Marchant et al., 2011), killer robots (Docherty, 2012, 2015), terminators (Garcia, 2014), and cyborg assassins (Garcia, 2014).

state's autonomous vehicle definition of driving "without… active physical control or monitoring" (Levandowki, 2016). Uber's argument has been characterized as "textually plausible but contextually untenable" as it only exploits a "linguistic loophole" in the California statute's definition of "autonomous technology" (Smith, 2016). After removing their autonomous vehicles from California in December, Uber has now applied for and received the necessary registrations and permits ("Testing of Autonomous Vehicles - State of California, Department of Motor Vehicles," 2017).

Cognitive engineering is not immune to the difficulties of defining human-automation systems. However, by virtue of its sociotechnical perspective, as laid out in Sec. II. Complexity, and its long history of examining such systems, cognitive engineering continues to develop and evolve relevant methods for characterizing them. In this section, we specifically examine the flaws in two of the most common types of definitional frameworks: levels of autonomy and human-in-the-loop. We will then conclude by providing a new approach for defining human-autonomous systems based on the distribution of work rather than the amount of automation or conceptual location of the human.

A. Problems with levels of automation and human-in-the-loop

To develop definitions of these human-autonomous systems, lawmakers and policymakers have typically relied on two increasingly outdated classification methods: levels of automation and human-in-the-loop. These two constructs are useful introductions to understanding human-automation interaction but they are entirely insufficient to formally define systems, especially with regulatory intent.

Levels of automation define the tradeoff in capability between a human and an automated agent within a system on either a single-dimensional scale (Billings, 1997; Sheridan & Verplank, 1978) or a multi-dimensional scale (R. Parasuraman, Sheridan, & Wickens, 2000). For example, the five levels of autonomy for highly automated vehicles developed by SAE International (SAE International, 2016) has been adopted as the standard vehicle categories within the U.S. National Highway Transportation Safety Administration's (NHTSA) Federal Automated Vehicles Policy (NHTSA, 2016).

Notionally, as the level of automation increases, the level of human involvement decreases; ultimately resulting in all actions being performed by the automation. This state of "no human involvement" is referred to as human-out-of-the-loop. All other states are referred to as human-in-the-loop

systems because there is some human involvement in the system operations. Whether humans are "in" or "out" of the loop has become the fundamental question of regulating autonomous weapons as many argue that without a human-in-the-loop, a weapon is inherently in violation of humanitarian and international law (Docherty, 2012, 2014, 2015; Ford & Jenks, 2016; Knuckey, 2014; UN CCW, 2016).

Despite the former prevalence of levels of automation in the academic literature, they are now acknowledged to be limited, problematic, and worth discarding altogether (Bradshaw et al., 2013; Dekker & Woods, 2002; Feigh & Pritchett, 2014; Lintern, 2012; Murphy & Shields, 2012). The general flaw of levels of automation is that they focus too much on a singular, static definition of the automation's capabilities, forgetting about what the corresponding human's capabilities will need to be. Examples of these flaws are present in the current definition of Level 2 Partial Driving Automation used to classify vehicles in the U.S., which in most states is the highest level of driving automation without being regulated as an "automated vehicle" (Smith et al., 2015):

> "The driving automation system (while engaged) performs part of the dynamic driving task by executing both the lateral and the longitudinal vehicle motion control subtasks, and disengages immediately upon driver request;
>
> "The human driver (at all times) performs the remainder of the [dynamic driving task] not performed by the driving automation system [e.g. object and event detection and response]; supervises the driving automation system and intervenes as necessary to maintain safe operation of the vehicle; determines whether/when engagement and disengagement of the driving automation system is appropriate; immediately performs the entire [dynamic driving task] whenever required or desired." (SAE International, 2016, p. 19)

Notice that the tasks directly related to driving are only described with respect to the automation, the lateral and longitudinal vehicle motion control e.g. lane centering, parking assist, and adaptive cruise control. The first stated role of the human driver is to perform "the remainder of the [dynamic driving task] not performed by the driving automation system." This is the definition of leftover allocation (Bailey, 1982): automate as many functions (or activities) as technology will permit, and assume the human will pick up whichever functions are leftover. Leftover allocation often results in humans being assigned the function of monitoring automation or the environment for

conditions beyond which the automation can operate (Bainbridge, 1983; Wiener & Curry, 1980); a function in which humans are ineffective (J. D. Lee & Moray, 1992; Molloy & Parasuraman, 1996). With these distributions of functions, workload will spike during off-nominal situations (Bainbridge, 1983) and be excessively low during normal operations between spikes, ultimately leading to humans who are "in-the-loop" becoming, practically, "out-of-the-loop" (Bainbridge, 1983; Endsley & Kiris, 1995). Despite these consistent findings against requiring human monitoring and supervisions, the Level 2 Autonomy definition requires humans to do exactly that.

The conclusion of the list of human functions for Level 2 Autonomy requires the human to determine "whether/when engagement and disengagement of the driving automation system is appropriate," and if disengagement is necessary, "immediately [perform] the entire [dynamic driving task]." Again, we must ask if these are appropriate functions to assign to humans in this context. In complex work environments where many functions are interdependent, coupled, and hidden, the driver is likely unable to determine when disengagement is "appropriate" (Javaux, 2002).[2] Studies have shown that unaccounted-for couplings can result in insufficient coordination, idling, and workload accumulation in human-automation interaction (Feigh & Pritchett, 2014). Furthermore, the SAE International and NHTSA standards require the human to "immediately" takeover control in off-nominal conditions but provide no discussion of how those emergencies should or could be supported – an example of brittle automation (Norman, 1990). Ultimately, the various levels of automation are acknowledged by automated car developers themselves as raising "particularly difficult issues of human-machine interaction which have not been satisfactorily solved" (Smith et al., 2015).

All this discussion of the ways that human-automation interaction can go wrong, shows how neither requiring a human in the loop, nor barring a human from the loop inherently makes a system more or less effective. Taking a complex system and simply requiring a human safety net could make the system less effective. On the other hand, meaningfully integrating a human into an automated system could make it perform much more effectively.

[2] Drivers in highly automated vehicles built according to the SAE International and NHTSA standards would be particularly unlikely to understand the nuances of the vehicles because ambiguity was built into the standards. The level assignment "expresses the design intention" such that performance deficiencies in the driving automation system does not automatically change the level assignment (SAE International, 2016, pp. 27-28). Even further, the standards state that a system can deliver multiple features at multiple different levels under varying conditions (SAE International, 2016, p. 28), expanding the number of vehicles states that the driver would have to account for.

Based a survey of legal cases involving automation, Jones (2015, p. 92 & 97) defined this reality as the "irony of automation law:"

> "When accidents happen or bias or abuse occurs, a mechanical fix is a tempting solution. Removing the human, this line of thinking goes, will remove the subjectivity, the errors, the inexactness, and the carelessness. This result is neither possible nor desirable and approaches automation as if it has had no negative consequences… Other times, machines appear to be the source of the problem. In order to quickly deal with the issue, the law has simply banned the lack of a human or required their involvement. Neither approach effectively solves the problem or protects the stated interests."

B. Toward work-based definitions of human-automation systems

If the levels of automation are too static and too focused on the automation's capabilities to be useful for defining human-automation systems and requiring a human in or out of the loop does not inherently relate to system effectiveness or safety, what framework should be used to help define and classify human-autonomous systems? We argue that definitions and classifications of human-automation systems should be based primarily on work, not the automation's capabilities or the presence of a human. As stated by Richards and Smart (2013, p. 21), "[designing] legislation based on the form of a robot, and not the function… would be a grave mistake." Therefore, completing specific work should be the requirement while the distribution of functions between the automation or human capabilities should be viewed as potential solutions. Said another way, if certain automation or human capabilities are necessary because of the work requirements, then incorporate them; if they aren't necessary, they can be excluded. In this subsection, we provide some insight into how work-based definitions could be developed.

Understanding the specific work demands and roles within a domain and allocating work appropriately to various agents is a fundamental question in cognitive engineering (see Sec. II. Complexity). Cognitive engineering views humans and automated systems as teammates and collaborators, who complete work together (e.g. a human driver and an automated vehicle, together safely driving to a destination). Based on a review of the literature, Feigh and Pritchett (2014) concluded that all human-automation teams, whether an individual level of automation for highly-automated vehicles or an individual deployment of autonomous weapon, ought to be meet five requirements for effective function allocation:

1. Each agent must be allocated functions that it is capable of performing.
2. Each agent must be capable of performing its collective set of functions.
3. The function allocation must be realizable with reasonable teamwork.
4. The function allocation must support the dynamics of the work.
5. The function allocation should be the result of deliberate design decisions.

Expanding on the perspective of teamwork and teammates, Johnson, et al., (2011; 2014) developed coactive design, a process that designs human-automated systems based on the identification and active management of the interdependence between the human and automated agents (Feltovich, Bradshaw, Clancey, & Johnson, 2006). Interdependence is the acknowledgement that what each agent does depends on what each other agent does; thus, requiring coordination in time and space, and some amount of transparency and trust (see Sec
IV. Transparency). Interdependence was used to design the winning robot at the 2013 DARPA Virtual Robotics Challenge (Johnson, Bradshaw, Feltovich, et al., 2014) and the outcomes of their design process read like definitions and classifications for regulating a specific human-automated system. Interdependence analysis identified the capacities required to complete the task, the enumeration of viable team roles, and the assessment of the capacity of the human or the robot to perform tasks or support each other. Their observability, predictability, and directability framework answered questions such as "What information needs to be shared," "Who needs to share with whom," and "When is it relevant."

These perspectives of teamwork between human and automated agents can form a useful foundation for policymakers and lawmakers. Their first step should be to identify the high-level work that needs to be completed e.g., the specific tasks, actions, or laws (see Cognitive Work Analysis in Sec. II-B). Then research programs and technical experts can use the models and measures of function allocation (Pritchett, Kim, & Feigh, 2014a, 2014b), independence analysis (Johnson et al., 2011; Johnson, Bradshaw, Feltovich, et al., 2014), and the like, to identify what sets of human-automation teams, technologies, and concepts of operation, are capable of adhering to the high-level work demands. From these sets, legal and ethical considerations can determine which sets best adhere to standards of transparency, safety, and accountability.

The most important characteristic of these work-based definitions is that they describe systems based on their ability to complete the work – essentially focusing on "what" should be done, rather than "who" does the work. The types of human-automation teamwork can span the spectrum from completely removing the human, to completely integrating the human into every function, to any combination in-between. If the overall goal of the human-automation system is to adhere to the laws of driving, the laws of war, or to a certain standard of safety, the work-based definitions do not overtly prescribe whether or where a human operator should be within the team, just that the laws and standards should be adhered to. This avoids the tendency to conflate what is legal or ethical with what is safe or effective.

One human-automation systems domain which has endorsed work-based definitions is commercial aviation and, in particular, the governance of the design and operation of flight guidance systems, including autopilot systems. The dramatic changes in technology and system design in recent decades has resulted in much higher levels of integration, automation, and complexity within the cockpit (Federal Aviation Administration, 2013). The 2014 FAA Circular for Approval of Flight Guidance Systems (Kaszycki, 2014) provides design guidance embodying the principles of effective function allocation and the perspective of interdependence to ensure the systems achieve their potential of better performance, increased safety, and decreased workload. The guidance uses the consistent convention of assigning primary authority for specific functions to specific agents without ambiguity but also lays out procedures for transitions of authority between the pilot and autopilot, and specifies the limitations, goals, and maximum acceptable error margins for the autopilot.

This type of design guidance and regulation should be the goal of discussions intending to define and classify certain categories of human-automation systems. Framing classifications of human-automation systems through work allows for the classifications to stay relevant as capabilities evolve and change. These work-based classifications avoid the focus on specific amount of automation or conceptual location of the human which can limit innovation and create loopholes in the regulations. By limiting the focus to completing the required work, these classifications provide space for other perspectives to constrain the allowable human-automation systems based on legal or ethical concerns, or as discussed in the sections to follow on transparency, safety, and accountability.

IV. TRANSPARENCY

> *Marvin trudged on down the corridor, still moaning. "And then of course I've got this terrible pain in all the diodes down my left hand side..."*
>
> *"No?" said Arthur grimly as he walked along beside him. "Really?"*
>
> *"Oh yes," said Marvin, "I mean I've asked for them to be replaced but no one ever listens."*
>
> -Douglas Adams (1997, p. 71)

In recent years, there has been an increasing call for transparency in automated and robotic systems (Castelvecchi, 2016; DARPA, 2016; IEEE, 2016a, 2016b; NSTC, 2016). The reasons given for this demand seemingly stem from three main concerns: trust, validation, and identifying the root cause of failure (DARPA, 2016; IEEE, 2016a; NSTC, 2016). However, transparency is generally vaguely defined (J. Y. Chen et al., 2014; Selkowitz, Lakhmani, Chen, & Boyce, 2015) and thus hard to measure (IEEE, 2016b; Owotoki & Mayer-Lindenberg, 2007). Previous work in human-robot interaction and cognitive engineering may provide grounded and validated methods to better define, resolve, and measure transparency.

In the late 1980's, researchers in psychology began to extend work on human-human trust to human-machine systems (Muir, 1987). They began to propose models of trust that included its sub-components, layers, and relation to levels of automation (Inagaki, Furukawa, & Itoh, 2005; Mirnig, Wintersberger, Sutter, & Ziegler, 2016; R. Parasuraman et al., 2000) and function allocation (Inagaki et al., 2005; J. D. Lee & Moray, 1992; R. Parasuraman et al., 2000). From these works, multiple definitions of transparency within trust crystallized, ranging from accountability to explainability and understandability to predictability. Many of these concepts seemingly overlap and a multitude of definitions and models were offered to incorporate them within a coherent framework (Goillau, Kelly, Boardman, & Jeannot, 2003a; J. D. Lee & See, 2004).

In these and other various models, trust is commonly decomposed into three levels: performance, process, and purpose, as synthesized by Lee and Moray (1992). These three levels generally map onto what the automation does, how it does it, and why it does so, respectively. While often understood in relation to trust, J. Y. Chen et al. (2014) have argued for this framework to extend to transparency for situation awareness. In their model 'how' and 'what' combine into 'Basic Information', while 'why' is

segregated into rationale and expected outcomes. Conversely, Marr (1982) combined the two upper levels of 'what' and 'why' into what he termed the computational level and instead split system operations into algorithmic and implementational components. This partition tended to separate hardware-based and software-based descriptions. Poggio (2012) later extended this explanatory model by adding two levels above the computational: learning and evolution. Finally, a fourth major tripartite system was proposed by Dennett (1989) from his philosophy of mind on attribution to agents. His system asserted that behavioral ascription can be predicated on a physical stance, a design stance, or an intentional stance. This means that humans attribute behavior either (1) based on physics, (2) based upon how a machine was designed to function, or (3) by ascribing beliefs and goals. Each of these approaches bring different models and nuances to the table but they ultimately can be synthesized into a six-level model of explanation or transparency, as presented in Table 1.

Table 1. Building a complete picture of transparency.

<table>
<tr><th></th><th>Lee and Moray</th><th>Chen</th><th>Marr + Poggio</th><th>Dennet</th><th>A Synthesized Model of Transparency</th></tr>
<tr><td>Why to decide what to do</td><td></td><td></td><td>Evolution</td><td></td><td>Evolution/Origin</td></tr>
<tr><td>How to decide what to do</td><td></td><td></td><td>Learning</td><td rowspan="2">Intentional</td><td>Learning</td></tr>
<tr><td>Why</td><td>Purpose</td><td>Rationale\ Outcomes</td><td rowspan="2">Computational</td><td>Purpose</td></tr>
<tr><td>What</td><td>Performance</td><td rowspan="3">Basic Information</td><td>Design</td><td>Performance</td></tr>
<tr><td>How (software)</td><td rowspan="2">Process</td><td>Algorithmic</td><td></td><td>Algorithmic</td></tr>
<tr><td>How (hardware)</td><td>Implementational</td><td>Physical</td><td>Implementational</td></tr>
</table>

Transparency for trust demands predictability and understanding, which, according to Zhu (2009), requires a purposefully-designed intentional stance when applied to complex, non-human agents. Unlike the other transparency models in Table 1, applying the intentional stance necessarily requires referential opacity in the Other's beliefs and desires (Foxall, 2005). In other words, at some level, the human must begin to think of the machine as an agent with internal states and not just a hunk of metal or lines of code. This approach is supported by Epley, Waytz, & Cacioppo (2007) as it encourages anthropomorphism of non-human agents, an attribution which increases explainability and may help build social trust (Breeman, 2012; Waytz, Heafner, & Epley, 2014).

Therefore, ascribing intentionality contains an essential contradiction, opacity is required for some level of understanding and therefore trust (Zhu, 2009). This concept is deeply rooted in Mayer et al.'s widely-accepted definition of trust as "the willingness to be vulnerable to the actions of another party" (1995, p. 712). Distrust is then, in part, expressed by monitoring and regulatory mechanisms (Lewicki, McAllister, & Bies, 1998), and transparency demanded to minimize uncertainty (Muethel & Hoegl, 2012). This tension in trust between transparency and opacity may be demonstrated by a simple analogy: a parent either trusts their teen or tracks their phone. Trust, then, as a social phenomenon, may require giving up some transparency; though developing a reputation for trustworthiness may require some degree of transparency to begin with. Thus, trust may evolve gradually moving from more to less transparent, from reliability to predictability to faith (Muir, 1987). It may also fluctuate as expectations and evidence for trust and distrust are integrated over the course of a relationship or interaction (Lewicki et al., 1998).

Trust, it behooves us to note, is not always good as it may be miscalibrated to over- or under-trust when inappropriate (J. D. Lee & See, 2004; Lewicki et al., 1998; Muir & Moray, 1996). Furthermore, studies have demonstrated that the more trust placed in the system the less situation awareness the human maintains (Goillau et al., 2003a) a phenomenon known as complacency (J. D. Lee & See, 2004; Raja Parasuraman & Manzey, 2010). Therefore, one may wish to reject the view that we should be trusting intelligent systems at all, let alone encouraging anthropomorphism or ascribing intentionality. This, though, is a perilous approach, as trust has been shown as key to the acceptance and use of such systems, whereas rejection can increase human workload and decrease safety (Desai, Stubbs, Steinfeld, & Yanco, 2009; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Floyd & Aha, 2016).

Helldin's thesis (2014) lays out a set of transparency guidelines for aircraft automation that echo our conclusions on transparency for trust. Beyond testing her model objectively, Helldin had operators rate each requirement's importance in performance subjectively and found that communication of purpose and feedback on learning and performance were perceived as most important. On the other hand, knowledge of actual rules, algorithms, raw data, or specific points of failure were ranked least important. These findings strongly support the framework, summarized below in Table 2, for trust in transparency. In line with Chen *et al.* (2015), Helldin (2014) also found that feedback on learning, purpose, and performance where helpful in identifying and closing the gaps between human and machine task

awareness. In her final requirement, Helldin (2014) urges design to be with an appropriate level of automation in mind. While levels of automation may not be the best framework for classifying systems, as discussed in Sec. III. Definitions, it underlines the importance of clear design guidelines that help inform the purpose of the system, help calibrate trust, and serve as a yardstick for validation.

Table 2. Transparency for trust, fault finding, and validation & verification.

A Synthesized Model of Transparency	Trust	Fault	Validation & Verification
Evolution/Origin	Stable	System Design	Stability
Learning	Intentional	Nested Query with Natural Termination	Convergence
Purpose			Design Guidelines
Performance	Robust, Predictable, Dependable, Reliable		Function, off-nominal, and anomaly detection
Algorithmic			Programming checklists, code reviews, traceability analyses, static analyses, and coverage analyses
Implementational		Diagnostic	Hardware Certification

At this point, one might argue that the type of transparency we need for assigning fault is fundamentally different from that needed for trust. Here, lessons may be drawn from cognitive psychology and neuroscience. What was found in these fields was that strides in explanatory mechanisms could be made decades in advance and have much higher explanatory power at an algorithmic or computational level than an implementational one (Epley et al., 2007; Vaina & Passingham, 2016). This makes sense; we would not necessarily want an expert witness to describe each bit passed through a sensor when describing why an autonomous car killed a pedestrian. Thus, there is a need for tailoring the level of transparency to the required context, even in fault assignment (Alaieri & Vellino, 2016; Langley, 2017). In this vein, work on explainable AI (XAI), in which the system has an expressive layer that can be queried and can customize appropriate answers, has begun (DARPA, 2016; Hill et al., 2006; Kay & Kummerfeld, 2012; Wang, 2012).

For XAI to prove sufficient for assigning liability or diagnosing fault, it will not only need to be able to match response appropriateness but also have the capability to nest its levels of explanation such that a line of inquiry can be followed to a logical yet reasonable termination. This does not mean that the system must always be able to reach the implementational/physical level but that it can recognize when there is no more sense in adding more detail, in other words having natural termination conditions. This is equivalent to a parent knowing when to stop responding after a child repeatedly asks "Why?" Just as we don't expect a human to be able to explain every action and its causes down to a cellular level, it may be useful to define the extent to which a system should be self-aware vs. when we will permit external diagnostics to diagnose fault or perform validation.

Validation of systems that learn and evolve is an open-ended problem, but one that cannot be ignored. While XAI aims to make underlying models derived from machine learning understandable, this only represents one part of the validation challenge. Jacklin et al. (2005) outlines other aspects of this multi-level process including static analysis for code, convergence analysis for learning, and learning algorithm stability analysis for how learning evolves. In later work, Jacklin (2008) laid out the gaps in knowledge and capability of current monitoring systems to perform learning stability analysis, off-nominal detection, and lack of clear guidelines from regulatory agencies.

To implement Jacklin's (2008) recommendations for validation and verification, there are many elements that must still be developed for transparency to be achieved. However, we are not left without guidance. Metrics for trust and system reliability have been developing over the past three decades (Desai et al., 2009; Dzindolet et al., 2003; Goillau et al., 2003a; Goillau, Kelly, Boardman, & Jeannot, 2003b; Hancock et al., 2011; J. D. Lee & See, 2004; Muir, 1987). Industry has already bought in to best programming practices, diagnostic fault and anomaly detection (Chandola, Banerjee, & Kumar, 2009; Spirkovska et al., 2010), off-nominal analysis (Belcastro, 2010, 2012), and hardware certification ("Directive 2007/46/EC of the European Parliment and of the Council," 2007, NHTSA, 2016, "The FAA and Industry Guide to Product Certification," 2004; NHTSA, 1999). Work on explainable AI is making headway in fault querying and that community has acknowledged the need for level-appropriate answers and nested explanations (Castellano, Fanelli, & Torsello, 2007; Lomas et al., 2012; Zhu, 2009). Further work must also be done on stability and convergence validation but a theoretical framework at least has been laid.

Finally, clear design guidelines for verifying system operation are imperative and has begun in realms such as autonomous vehicles ("Chapter 482a - Autonomous Vehicles," 2013, "Deployment of Autonomous Vehicles for Public Operation - California DMV," 2017, NHTSA, 2016) and UAVs (FAA, 2016; Haddon & Whittaker, 2003) but must be expanded and generalized to other intelligent systems.

V. ACCOUNTABILITY

"Accountability is emphasized… because the way decision makers are held accountable is presumed to influence how they make those decisions and the quality of those decisions"

– David Woods

It is desirable to be able to predict or prevent accidents in complex sociotechnical environments with human-automation systems. However, accidents will happen, and identifying the underlying factors and the interrelationships among the actors accountable for the accident are fundamental legal issues (Grosz et al., 2016; IEEE, 2016a). Often the primary concern of these interactions is the apparent ambiguity of who should be responsible should system failure result in, among other things, loss of money, property, or life. This responsibility gap (Müller, 2016) might result in the operators, programmers, or manufacturers escaping liability (Docherty, 2012, 2014, 2015). At the very least, the safety analysis and the lawsuits for human-automation system accidents will be expensive with outcomes that are hard to predict, potentially resulting in recalls (Greenblatt, 2016).

Much of the recent attention on accountability within human-automation systems has been focused on the perspectives of computer science and privacy law. These are important collaborations between legal scholars, ethical scholars, and technologists. However, we stress that just as automation is only one of nine factors of complexity in sociotechnical systems (Sec. II. Complexity), ensuring that only algorithms are accountable is insufficient for making the complete human-automated system accountable. All nine factors of complexity are becoming more commonplace, making the ideal of eliminating accountability gaps an increasingly difficult task.

In cognitive engineering, there are methods for modeling the factors and relationships for accountability at both the early stages of design as well as after an accident has occurred. Specifically, there are two broad areas where cognitive engineering can be useful in narrowing the accountability gap. The

first area is in effective function allocation in human automation teams where agents are allocated functions with clear expectations in terms of authority and responsibility. The second area concerns analysis of adverse events in complex sociotechnical systems that helps identify why complex systems fail and the contributions of various actors involved.

A. Accountability in early-stage design

Function allocation is a design decision that determines how agents (human or automated) will share workload and who will be held accountable if an adverse event occurred during the operation of the system. Function allocation pertains to the division of taskwork within a human-automation team and the actions involved in coordinating the teamwork to support taskwork division. Function allocation involves both 'authority' and 'responsibility' assignments. Authority implies an agent is designated to execute the action. Responsibility implies an agent has the accountability for its outcome in a legal or regulatory sense. The notion of authority and responsibility can be thought of as a mapping from the set of agents to the set of functions constituting the concept of operation.

Feigh & Pritchett (2014) provide a critical review of function allocation, leading to a set of universal requirements that any function allocation should satisfy. These requirements in turn, result in guidelines and a language to inform the design of human-automation systems. Responsibility pertains to who is held accountable for the outcome of a given task within an organizational or legal context. An agent that is responsible for a task is not always the agent who is assigned to execute the task, known as the responsibility-authority double bind (Feigh, Dorneich, & Hayes, 2012; Woods, 1985). Function allocation is a critical design decision concerning the allocation of work functions to human and automation agents in a human-automation team. The decision impacts the breakdown of the taskwork involved in accomplishing work goals as well as the teamwork necessary to coordinate the execution the allocated functions between agents in the team.

Take, for example, the task of driving a car from point X to point Y in a city without an electronic navigational aid. Now consider two functions – performing navigational decisions and controlling the car's speed and direction. Consider two possibilities in function allocation: (a) The driver has authority for and is responsible for both functions as she knows the city map well; (b) The driver has responsibilities for both functions however, a passenger is given authority to perform navigational functions as the driver does not know the city map well but the passenger does. This division of

taskwork in '(b)' would require the passenger and driver to communicate about the directions to successfully arrive at point Y. In '(b)' the work being performed by the driver to control the car's speed and direction is an example of taskwork. Similarly, navigational decisions about the car's route and direction being made by the passenger is another example of taskwork. The overhead resulting from communication between the driver and passenger is what we term teamwork.

A synthesis of the function allocation literature makes a clear argument for a general design principle: "the responsibility for the outcome of a function must be considered relative to the authority to perform it" (Pritchett, Kim, & Feigh, 2014a). In other words, when referring to an agent being responsible for a given function it is also important to state who has the authority to execute that function. Failure to follow this principle can result in authority-responsibility mismatches: automation is assigned to execute a function in an operational sense but the human will be held accountable in an organizational and legal sense for its outcome (Woods, 1985). Without being able to assess whether automation is correct, humans often overtrust or undertrust the automation (Raja Parasuraman & Riley, 1997).

An example of a well-studied mismatch in the context of modern commercial airline cockpits is provided by Pritchett et al. (2014a). Within this example, human flight crews have the responsibility for maintaining flight safety while autopilot and autoflight systems have significant authority over critical functions of aircraft control and trajectory management.

Note that the number of mismatches can be both desirable or undesirable depending on the context in which the sociotechnical system operates. Mismatches could be undesirable when minimizing total information transfer between agents is sought after, reflecting a situation where concerns about communication bandwidth, cyber-security, or failures in information transfer are paramount. Mismatches could be desirable when maximizing total information transfer between the clusters, reflecting a desire for maximum redundancy and error checking through shared situation awareness or cross-checking between agents.

Pritchett et al (2014) critically reviewed existing methods of work analysis such as Cognitive Work Analysis (Rasmussen et al., 1994) and Contextual Design (Beyer and Holtzblatt, 1998). They identified these methods' shortcomings especially from the standpoint of assessing the accountability vested in the agents operating the system. In response to these drawbacks, and motivated by the universal requirements of function

allocation, Pritchett et al. (2014a; 2014b) developed conceptual models of various function allocations between the human flight crew and the autopilot systems and counted the number of mismatches. Then, to understand the effect of the various function allocations within different contexts, the models involved computational simulations to assess the function allocation in terms of metrics (Pritchett et al, 2014b).

Computational models were developed to simulate and evaluate function allocations to inform the design process of an automation system. This evaluation is based on five requirements of human-automation teams, as identified in Pritchett et al. (2014b). As described in Sec. III. Definitions, the levels of automation framework is not adequate for describing dynamic interactions between the functions in a human-automation team. Work Models that Compute (WMC) is a computational model that can simulate the dynamics of the interactions between human and automated agents over time (Pritchett et al., 2011). Simulations are performed on specific scenarios where each one defines the operational environment of the human-automation system. In addition, the simulation can generate a number of measures for function allocation including incoherency, workload, work environment stability, authority-responsibility mismatches, interruptive automation, among others (Pritchett et al., 2014a). The dynamic nature of the computational model is particularly useful in predicting spikes in workload at specific points in time within a given scenario.

WMC can be used to model the above example of two functions involved in travelling between points X and Y with a car. Computational comparisons can be made between the function allocation decisions of '(a)' and '(b)' in terms of coherency and other measures. The assessment of coherency of the function allocation could assist designers make decisions about 'functional blocks' required for a given human-automation team (Bhattacharyya, 2016). For example, it may be determined that the functions of navigational decision-making and speed and direction control should be combined into a single block and thereby best performed by a single agent.

Accountability for the outcome of functions engenders the need for monitoring in complex systems wherein the agent with responsibility must monitor, and maybe intervene in, the actions of the agent with authority to perform a certain function. WMC has been used to demonstrate such monitoring for air traffic concept of operations (Pritchett and Bhattacharyya, 2016). While the focus of that study was on the monitoring requirements inherent within a range of function allocations, it also highlighted a spectrum of monitoring behaviors that arise from accountability. These ranged from

basic monitoring, where the agent with responsibility confirms the correct execution of the function, to complete monitoring, where the agent with responsibility intervenes and takes over the execution of the function.

By using WMC in the early stages of design, the impact of accountability can be assessed through computational modeling and simulation to pinpoint authority-responsibility mismatches, coherency of the division of tasks between agents and potential monitoring overheads on the agents. All of which enable a quantitative discussion of accountability, which can guide the design and development of the system and the concept of operations going forward.

B. Accountability after accidents

In cognitive engineering and safety science, Rasmussen's (1997) risk management framework has been applied in several in-depth analyses of large-scale accidents. They include such diverse events as the Walkerton E. Coli outbreak in Canada (Vicente & Christoffersen, 2006), the Flash Crash of May 6 2010 in US financial markets (Minotra & Burns, 2016), the alarming rate of mishaps in road freight transportation in the US (Newnam & Goode, 2015), the spread of beef contamination in the UK (Cassano-Piche, Vicente, & Jamieson, 2009), the Sewol ferry accident in South Korea (Kee, Jun, Waterson, & Haslam, 2017), and the police shooting of an innocent man in South London (Jenkins, Salmon, Stanton, & Walker, 2010).

The risk management framework is a product of four decades of research on risk management of complex sociotechnical systems (Vicente & Christoffersen, 2006). The underpinnings of the framework can be briefly described as follows:

- It describes sociotechnical systems as several interconnected levels that require 'vertical integration'. These levels from lowest to highest are approximately as follows – physical process layer, staff workers, management, regulators, and elected government law-makers. Decisions at the top-level should propagate downward and activities at the lowest level should be visible at higher levels, collectively referred to as 'vertical integration'
- The theory posits that safety can be jeopardized by a loss of control associated with insufficient vertical integration between the levels of a sociotechnical system described above.

- In a sociotechnical system that is about to release an accident, the behavior of actors change over time; such changes are adaptations to economic and psychological (i.e. cognitive load) pressures, and happen at multiple levels. Several such adaptations may move the sociotechnical system closer to the boundary of safe performance in a gradual manner. These behavior changes are referred to as migrations in work practices.
- When a sociotechnical system is at the boundary of safe performance, large-scale accidents are caused when some unique catalyst 'releases' the accident.
- Large-scale accidents are not caused by any single one-time threat to safety but by several migrations in work practices over a period of time followed by a catalyst.

Why should the framework be adopted in the legal community? As the framework explains large-scale accidents, its underpinnings offer guidance to analyze large-scale accidents and to identify interrelationships between various entities involved. The theoretical framework provides the analyst with concepts pertaining to large-scale accidents; for any particular large-scale accident, it helps abstract distinctive details of that accident and derive generalizable findings and high-level patterns (Vicente & Christoffersen, 2006). These generalizable findings can be used towards making decisions to prevent similar accidents from occurring. An associated technique, AcciMap, can be used in the analysis to display relationships in a diagram and communicate about the results of the analysis succinctly in a multi-disciplinary team (Jenkins et al., 2010; Kee et al., 2017). However, accident analysts are sometimes interested in identifying accountable actors and their relationships in the context of the events leading to the accident. According to a study by Kee et al. (2017) on the South Korean Sewol ferry accident that took the lives of over 300 passengers, Rasmussen's framework and the AcciMap technique provide a broader picture of an accident, balancing individual accountability with the systemic factors associated with systematic migrations in work practices over time. In the analysis of an accident with Rasmussen's framework, all levels in a sociotechnical system should be given attention. Considering relationships among the various actors across all relevant levels of the sociotechnical system can avoid oversimplification of the accident, hindsight biases, and unfair blame (Kee et al., 2017). An example of a relationship that needs to be considered in identifying accountable actors can be the socio-political pressure faced by an inspector as a result of conflicts of interest arising from the inspector's coworkers (Kee

et al., 2017).

Rasmussen's risk management framework is comprehensive and it considers the social and organizational factors underlying an accident. Risk assessment methods that can be used to assess complex systems like nuclear power plants do not consider these factors. Reason's Swiss cheese model (Reason, 2000) has similarities with Rasmussen's framework as it considers 'latent factors' and 'active failures', however, it is not as comprehensive as Rasmussen's framework that provides several propositions some of which pertain to 'latent factors'. The systems-theoretic accident modeling and processes (STAMP) is another approach that can be used to analyze an accident and propose changes in control mechanisms to prevent future accidents in a given system. STAMP has been inspired by Rasmussen's framework however, its taxonomy of control failures is not flexible for identifying failures in social and organizational aspects of a complex system (Salmon et al., 2012).

Case Study: 2010 Flash Crash Examined with the Risk Management Framework

As the risk management framework is mainly motivated by the fast pace of technological adaptation in today's organizations, the analysis of the Flash Crash in US financial markets with the framework (Minotra & Burns, 2016) arguably offers a unique example of the framework's ability to explain accidents involving systems consisting of humans and automation.

The Flash Crash was an event in US financial markets in May 6 2010, where market prices on several financial products plummeted by over 10% and reversed within a few minutes on the same trading day. Additionally, it was during this period during which several financial products were traded at more extreme prices (e.g. a penny for a share or $100,000 per share). Numerous investors suffered losses if they placed orders at inopportune moments during this event. The plummet in prices started at the Chicago Mercantile Exchange (CME) and this propagated across several markets. The initial plummet at the CME was associated with high volatility, an order book imbalance and high selling pressure. Algorithmic trading systems were an integral part of the decisions made by market participants leading to the crash.

Rasmussen's framework helps tie together multiple sources of evidence pertaining to the factors involved in a given adverse event, the lack of vertical integration in the sociotechnical system, and the change in behavior or work practices over time. In the Flash Crash, the multiple factors involved were,

high market volatility, unusually high selling pressure from Waddell & Reed, the activity of high-frequency traders, stub-quote usage, inadequate transparency on regulatory criteria for breaking trades, cross-market arbitrage, the widespread use of algorithmic trading, Navinder Sarao's manipulative algorithm, and the potentially inadequate market surveillance (Minotra & Burns, 2016). The multiple factors are associated with different actors and some of these factors are intermittent while other factors pertain to inadequate vertical integration coupled with migrations in work practices over time. While Sarao's manipulative algorithm was possibly the only illegal activity that played a role in this event, other actors also contributed to the magnitude of the price declines and the propagation of the declines across multiple markets.

On the day of the Flash Crash, the lack of vertical integration was evident in inadequate transparency pertaining to regulatory criteria for breaking trades which possibly resulted in a number of market participants to withdraw liquidity owing to uncertainty associated with breaking trades (Madhavan, 2012). Insufficient vertical integration also pertains to inadequate feedback received by regulators - the insufficient surveillance present on May 6 2010, may have given way to market participants like Navinder Sarao who monetarily benefited from illegal market manipulations; this illegal activity surfaced many years later.

Recent advances in financial technology including algorithmic trading and high-frequency trading have given rise to more challenges in financial regulation. Advancements in financial technology allow traders to conduct over 100,000 transactions per second (Buchanan, 2015). While high-frequency traders may provide liquidity in financial markets, aggressive trading activity combined with a substantial presence of high-frequency traders could result in unfavorable events like the Flash Crash which took only a few minutes to develop on May 2010. The relatively slow pace of adaptation in regulatory policy and technology yields such adverse events. In other words, if there is sufficient vertical integration including transparency in regulation and surveillance technology to detect malicious activity sooner, triggers akin to the ones underlying the Flash Crash would have a smaller impact on the overall financial market. Rasmussen's framework provides this language to identify the relationships between actors underlying an adverse event that has occurred, and how the system may need to change in order to improve the safety of the system. It is unlikely that change in any one sub-system or level in the sociotechnical system may improve overall safety as we have seen that multiple actors and several systematic migrations in work practices are involved in a large-scale event like the Flash Crash.

Rasmussen's framework and the AcciMap technique have their limitations wherein they are only adequate in explaining accidents. We are aware of the valuable insights the framework can provide about a sociotechnical system in which accidents have occurred. However, the framework does not offer guidance to develop new organizational structures for complex systems involving complex automation or to evaluate existing ones. Revising Rasmussen's framework and incorporating insights and principles from other frameworks like WMC may bring about new capabilities to help evaluate sociotechnical systems. Implementing the approach, especially in accident analysis and system evaluation, would provide academics with more insights and feedback about its strengths and weaknesses in making sociotechnical systems safer. As Vicente & Christoffersen (2006) appropriately put it, "To evaluate whether the framework is indeed capable of enhancing safety in this way, government and corporate policy makers would have to adopt this approach to design or redesign the growing number of increasingly complex technological systems that surround us" (p. 110).

VI. SAFETY

> *"Accidents that result in dramatic casualty tolls such as the Chernobyl, Bhopal, or Piper Alpha tragedies, or accidents that cause significant environment environmental damage, such as the Exxon Valdez oil spill, or accidents that challenge national symbols of technological prowess such as the loss of two space shuttles Challenger and Columbia and their crews, are stark reminders that safety vigilance should always accompany technological endeavors, and that complacency in the design and management of complex sociotechnical systems will inevitably compromise safety and create latent conditions that can trigger accidents." – Saleh, Marais, Bakolas, & Cowlagi* (2010, p. 1106)

Safety concerns are one of the largest motivating factors for governing complex, sociotechnical systems, especially human-automated systems[3]. Prior sections have listed many of the concerns: the numerous factors contributing to the complexity of human-automated systems, the difficulty in defining and classifying these systems, the difficulty of achieving appropriate levels of transparency and trust, and the mismatch between responsibility and

[3] While other concerns, such as security and privacy, have different consequences, the problems and solutions addressed in the system safety literature are likely sufficiently analogous for the reader in non-safety critical domains to find some use in the material.

accountability. The rise of complex, sociotechnical systems have also created new problems for the researchers of reliability engineering, system safety, and accident causation (Felder & Collopy, 2012; Reason, 1997; Sheridan & Parasuraman, 2005).

This section provides a primer on how to think about the attributes and causes of system safety and accident causation. These domains of reliability engineering, system safety, and accident causation have by no means, "solved" the questions of how to define to build a 'safe system' or, even, measure "safety;" nor is there comprehensive agreement on the way forward, but we hope that it finds an audience in the legal community. The goals this section are the following:

1. To situate the preceding discussions in this work within real world context and to give concrete examples of how and why human-automated systems fail.
2. To show how tools and frameworks from the safety community can be utilized to proactively anticipate future failure modes and mechanisms.
3. To explain how quantitative measures of risk may not be applicable to human-automated systems with emergent features.

The remainder of this section is organized as follows. First, we discuss what methods are in place to prevents accidents from occurring, and how can those methods can fail. Then we conclude with a discussion of the popular safety analysis method, quantitative risk assessment (QRA), and how the emergent features of human-automated systems limit the usefulness of QRA and make human-automated systems accidents often rare and severe.

A. Barriers and contributors to accidents

One of the most pervasive safety principles within the system safety literature is called Defense-in-Depth.[4] The three primary goals of Defense-in-Depth is place safety barriers designed to 1) prevent incidents or accident initiating events from occurring, 2) prevent these incidents or accident initiators from escalating should the first barriers fail, and 3) mitigate or contain the consequences of accidents should they occur (J.H. Saleh et al., 2010; Sorensen, Apostolakis, Kress, & Powers, 1999). Although it was developed by the Nuclear Regulatory Commission, Defense-in-Depth

[4] Defense-in-Depth was developed originally by the Nuclear Regulatory Commission but exists in other industries with different names such as layers of protection (Summers, 2003).

functions more as a way to think about a problem, as opposed to giving concrete regulatory guidance (Sorensen et al., 1999). For a technical discussion of how to represent safety barriers in formal ways see (Duijm, 2008; Harms-Ringdahl, 2009; Sklet, 2006) and for a more holistic discussion about the formal methods and how they fit within the context of a sociotechnical system, see Hollnagel (2016).

Components of an accident and the safety barriers intended to stop their propagation can manifest at many different locations throughout a sociotechnical system. Saleh and Pendley's (2012) safety levers framework provides a comprehensive perspective of six sets of stakeholders who can both contribute to accidents or contribute to defense against accidents. As shown in Table 3, the six levers include everyone who interacts with the system, from regulators to operators, and from designers to researchers and educators. Each of these stakeholders may play critical roles in the safety of these complex, sociotechnical systems ways, but their interactions are often very nuanced. For a discussion of the nuances of these levels, see Reason (1997).

Table 3. Safety lever and stakeholders.

Safety lever	Stakeholder
Regulatory	Accident Investigators, safety inspectors, and regulators
Economic	Insurers (penalties) and Shareholders (incentives)
Organizational Managerial	Managers and company executives
Operational / Maintenance	Technicians and operators
Technical Design	Engineers and system designers
Research & Education	Researchers, academics, and students.

Below we elaborate on the events leading to death of 346 people aboard the DC-10 Turkish Airlines Flight 981 as an example of how every system accident has multiple safety levers and stakeholders which contribute to the accident. As an aside, we hope that this provides a reason to be skeptical of calls to develop methods for guaranteeing legal accountability in human-automation systems (Docherty, 2015; IEEE, 2016a). For Turkish Airlines Flight 981, and so many other system accidents, guaranteed accountability is an impossibility that can be approached but never fully achieved. Throughout the narrative of Turkish Airlines Flight 981 below (Fielder & Birsch, 1992; Fielder, 1992; Vesilind, 2001), we highlight the various safety levels contributing to the accident in brackets.

On March 3, 1974, Turkish Airlines Flight 981, a DC-10 passenger aircraft, experienced catastrophic decompression outside of Paris resulting in the deaths of all 346 on board. The cause of the catastrophic decompression is largely attributed to the flawed design of the cargo doors (Fielder, 1992). The design allowed the manual handle to latch the door in the shut position and signal to the cockpit that the door was effectively locked, but without the lock pins being engaged (the main source of strength) [Design]. The DC-10 passed its tests and was deemed airworthy and certified by the U.S. Federal Aviation Administration (FAA) [Regulatory]. On June 12, 1972, the rear cargo door of a DC-10, American Airlines Flight 96, blew off while flying over Windsor, Ontario. Luckily no one died as the pilots were able to maintain control. An investigation by the U.S. National Transportation Safety Board (NTSB) concluded that the loss of the rear cargo door caused the catastrophic decompression and ordered modifications to the locking mechanism of cargo door to prevent similar accidents. The NTSB recommended that the FAA ground all DC-10s until such modifications were made. However, due to the potential economic issues for McDonnell Douglas, the DC-10's manufacturer, the administrator of the FAA made a "gentleman's agreement" with the CEO of McDonnel Douglas that McDonnell Douglas would get the problem fixed without formal requirements or public notices [Economic/Regulatory]. McDonnel Douglas decided that instead of fixing the latch and lock pin mechanism, they would add a small window on the door to view the lock pins with instructions to cargo handlers to visually ensure the locking mechanism was in place [Design].

Within two weeks of the so-called "Windsor incident," Dan Applegate, the chief product engineer at Convair, the subcontractor in charge of designing the cargo door, wrote a memo to his management explaining that the next catastrophic loss of decompression would likely cause a loss of the airplane (Eddy, Potter, & Page, 1976). Trusting the FAA certification, and not wanting to lose the contract with McDonnel Douglas, or open themselves up to criticism or liability, the Convair management quieted Applegate and ignored the memo [Organizational – managerial]. Applegate did not blow the whistle and no action was taken [Education]. It did not take long for the NTSB's and Applegate's warning to become reality. Within two years, Turkish Airlines Flight 981 took to the air with unengaged lock pins on the cargo door, experienced catastrophic decompression at

> 23,000 ft., killing all 346 people on board. As it turns out, the Turkish baggage handlers had trouble shutting the door, but did not look through the added window because the instructions were written in English which they could not read [Operational – maintenance].

B. The seductive call of quantifiable risk for human-automation systems

Defense-in-depth and safety barriers provide a useful philosophy for conceptualizing accident prevention and containment; however, it does not provide a method for prioritizing certain barriers over others. One of the most prevalent proposed solutions is quantitative risk assessment (QRA), also known as probabilistic risk assessment (PRA). The mathematical and statistical basis for risk has been immensely popular and has been used in many successful applications. The most commonly used methods are the event tree analysis and fault tree analysis for identifying causal relationships (Rausand & Hoyland, 2003). Event tree analysis is the inductive method asking "what happens if?" whereas fault tree analysis is the deductive method, asking "what can cause this?" Through probabilities and structures of each cause, event, and branch in the trees, specific probabilities can be assigned to individual scenarios.

The numerical representations of risk can be seductive, but it is essential to acknowledge its flaws. First, fault trees are only as good as the information used to build them. Accidents can often occur from a failure to imagine failure. Additionally, probabilities can be based on beliefs about the world (Bayesian, do we expect the system to fail?) or based on prior data (Frequentist, has the system ever failed before?), and in some circumstances, these different probabilities lead to different conclusions. Using the correct set of data and techniques is essential to accurate calculations of the trees.

Second, there is no generally applicable method for comparing the risk of a given hazard against another (Stanley Kaplan & Garrick, 1981). Risk is generally defined as triplet (not a scalar) of the three following questions: 1) What are all the ways the system can fail (scenarios)? 2) How likely are those scenarios to occur? 3) How severe will the consequences be should the scenario occur? Integrating these three risk factors is dependent on the individual domain and perspectives of the stakeholder.

Third, and more importantly for this paper, QRA is limited when applied to software errors and sociotechnical issues, specifically with respect to accounting for rare or emergent events. QRA ultimately grew out of the

quality control of physical manufactured components, and thus it is easy to see why it has been so well suited for predicting accidents that stem from mechanical failures. Software and sociotechnical systems, however, are characterized by a significantly different failure mechanism: emergence.

Emergence occurs when interaction among components within a system (e.g. human or automated agents, technology, organizations) interact to cause outputs or states which cannot be predicted by accounting for the deterministic behavior of the components (Baldwin, Felder, & Sauser, 2011, p. 306; Felder & Collopy, 2012, p. 320). We focus on this issue because emergence is a natural trait of complex, sociotechnical systems, including human-automated systems (Baldwin et al., 2011; J. H. Miller & Page, 2007). Emergence has two main implications for human-automated systems that have been hinted at throughout this section and throughout this paper, but which are restated here: human-automation interaction errors are often emergent and traditional methods of statistics and experimentation often cannot predict them.

Humans obviously have their own emergent behaviors. But with respect to automation, and particularly within aviation, it has been shown that "instead of eliminating error, automation… generates new types of error arising from problematic human-automation interaction" (Pritchett, 2009, p. 83). For more examples see Sheridan and Parasuraman (2005), Wiener & Curry (1980). As discussed in Sec. II, III, and V, to address these emergent human-automation interaction issues, many researchers are attempting to better model and measure human-automation interaction through function allocation (Pritchett et al., 2014a, 2014b), cognitive work analysis (Vicente, 1999), interdependence (Johnson, Bradshaw, Feltovich, et al., 2014) and others. Similarly, there has been an increased emphasis for incorporating standards and minimal specifications of human-automation interaction (Courteney, 1999; "Flight Deck Alerting System (FAS)," 2007).

Despite the increased focus on understanding, modeling, and simulating human-automation interaction, there are still serious obstacles to be overcome. Here are only a few: Creating naturalistic test environments to examine rare and hazardous human-automation interaction situations is quite difficult, and humans are quite adept at averting abnormalities before they turn into the desired test cases (Pritchett, 2009). Computationally simulations have to account for an extraordinary range of interactions in which the events are rare and systematic recorded data even rarer, making duplicating exact circumstances difficult (Klyde, McRuer, & Myers, 1995; McRuer, 1995).

With respect to mathematical representation, there are challenges in even quantifying or statistically describing emergent behavior. As summarized by Felder & Collopy (2012, p. 321), "we are building systems that spend more time in nominal operation (that is, are generally better behaved) than previous generations, but when they do operate off nominal, are much further from the nominal than previous generations." This attribute of complex systems is related to the probability distribution of an off-nominal (likely emergent) event occurring. In traditional systems, the traditional statistics and reliability measures work because the probability distribution is normal (six sigma and most standard statistics measures work). However, as shown in Figure 3, in complex systems, the distribution is highly leptokurtic: that is, it is much "spikier" with the probability of normal behavior higher than expected, but the probability of off-nominal is much higher (fat tails) (Felder & Collopy, 2012, p. 321). Therefore, while new automated systems may hold the promise of reduced likelihood of failure in a general sense, they tend to increase the likelihood of more severe failures.

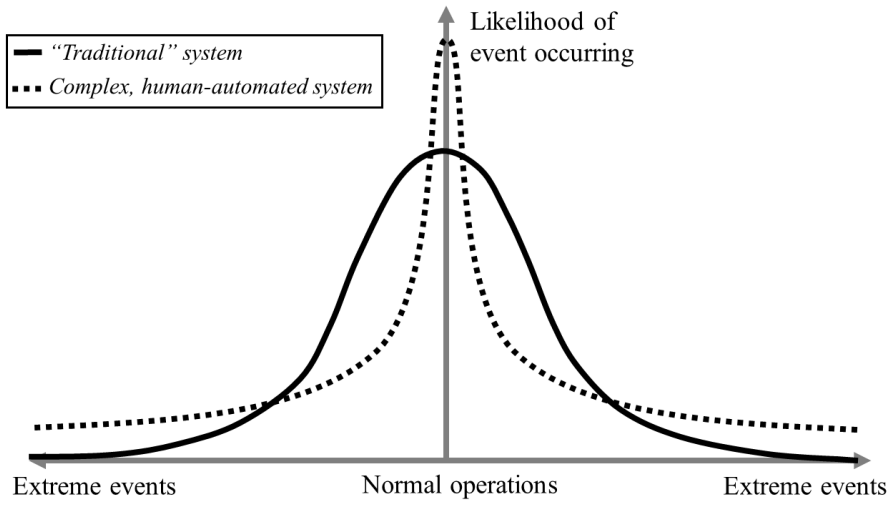

Figure 3. Notional statistical description of how more complex systems tend to increase likelihood of normal operations but also increase the likelihood of extreme events.

To provide a specific example of how human-automated systems can possess this duality of more normal operations with more off-nominal events, we examine the situation of piloted control of aircraft. First, the specific mathematical representations of aircraft handling qualities have been less-than-successful in characterizing pilot-aircraft interaction. Second, pilot loss-of-control and controlled flight into terrain are very rare occurrences in modern commercial aviation, but when they occur, are almost always catastrophic.

One of the major causes of loss-of-control in-flight (LOC-I) is aircraft-pilot coupling (or more commonly pilot-induced oscillation[5]). There are three levels of aircraft-pilot coupling. The first and second levels are based purely on mathematical measurements of the control algorithms: traditional linear dynamics (Level 1) and classical non-linear dynamics (Level 2). When developing the categories, the developers of the levels were forced to take into account emergence and thus constructed an "other" category for these human-automation interaction issues: "other non-linear dynamics" (Level 3) (McRuer, 1995). In remarks about Level 3, authors noted that the dynamics of Level 3 are substantially opaque with possibilities that are "difficult to identify or discover without an elaborate search" and that there exists "no universally applicable criteria" for its definition (Klyde et al., 1995, p. 98).

This issue with "other" or emergence is not just a mathematical issue, but ultimately, one of the most critical safety issues of modern aviation. Loss-of-control in-flight (LOC-I) and controlled flight into terrain (CFIT) are the first and second leading cause of fatal accidents in air transportation worldwide (Figure 4) (IATA, 2015). That LOC-I and CFIT receive "substantial industry attention despite a relatively low number of accidents is due to the disturbing number of fatalities they have produced" (IATA, 2015, p. 5). As can be seen in Figure 4, for other major high-risk accidents like mid-air collisions, we have either eliminated their likelihood or reduced their fatality rate. But what remains are rare, off-nominal events related to the control of the airplane, in other words, human-automation interaction.

In conclusion, human-automation systems can be summed up in the following way: They don't fail often, but when they do they fail surprisingly and spectacularly (Felder & Collopy, 2012).

[5]Given the increased understanding of emergence within complex aviation control systems, new terms have been introduced to replace the familiar PIO such that the pilot's guilt in such events is less likely to be assumed, including aircraft-pilot coupling (APC), pilot–in-the-loop oscillations and pilot-assisted (or augmented) oscillations (Klyde et al., 1995) (Witte, 2004).

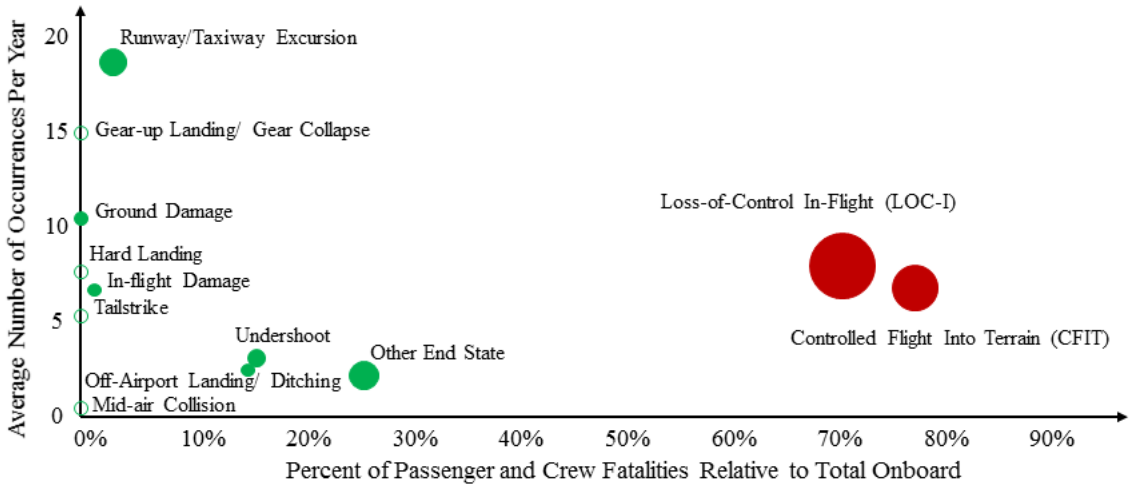

Figure 4. High-risk accidents in commercial aviation worldwide from 2010 to 2014 by frequency (average number of occurances per year) and fatality rate (percent of passenger and crew fatalities relative to total onboard when accident occurs). Bubble size increases as the absolute number of fatalitites for that category increase. Loss-of-control in-flight (LOCI) and controlled flight into terrain (CFIT) caused 1,242 and 707 fatalities, respectively. Figure adapted from (IATA, 2015, p. 5).

VII. Summary and Conclusions

A. Summary

The immense literature, perspectives, models, and measures, synthesized and introduced in this paper describe the strong foundation that cognitive engineering and its related disciplines can provide for governing human-automated systems in addressing the five main questions summarized below.

1. Complexity: What makes human-automation systems complex?

There are numerous attributes that contribute to the complexity of human-automated systems from both sociotechnical and naturalistic perspectives, most notably the difficulty of understanding human-automation interaction (Orasanu & Connolly, 1993; Vicente, 1999). Dealing with complexity becomes even more challenging when considering future systems because our visions of the future are often too specific, too vague, ungrounded, or overconfident (Woods & Dekker, 2000). The cognitive engineering

community has dealt with these complexities of human-automated systems for many years, developing frameworks such as cognitive work analysis (CWA, Vicente, 1999) and naturalistic decision making (Orasanu & Connolly, 1993) to better characterize and understand them. CWA seems to have particular use for legal frameworks as CWA focuses on defining the constraints on a system and how those constraints affect system performance. In the face of all the complexity of human-automated systems there are two rules that lawyers can take away: first, be cognizant of both the physical aspects of these systems (the design and arrangement of humans and automated systems) and problem-based aspects of these systems (the cognitive demands on the humans). Second, technology and automation are only hypothesized solutions for completing the work safer and more effectively, and thus should be evaluated based on how well they support that work.

2. Definitions: How should we define and classify different types of human-autonomous systems?

Definitions are the foundation for regulations and governance. They are also a major source of debate, confusion, and loopholes that often stymie substantive discussions about other important questions such as transparency, safety, and accountability. Two common types of definitions, levels of autonomy and human-in-the loop, which are prevalent in discussions of highly-automated vehicles and autonomous weapons systems were shown to be too static and too focused on the automations' capabilities to be useful for defining and classifying human-automated vehicles (Bradshaw et al., 2013; Dekker & Woods, 2002; Feigh & Pritchett, 2014). Designing human-automated systems based on these perspectives has often left humans responsible for monitoring and taking over for the automation when there are conditions beyond which the automation can operate. This design philosophy typically results in excessively low workload during normal operations punctuated by excessively high workload during off-nominal situations due to the interdependent, coupled, and hidden complexities. To address the problems with levels of autonomy and human-in-the-loop, we argued that definitions and classifications should be based on the system's ability to complete work, not the automation's capabilities or presence of a human. Evaluating two major work-based design perspectives – requirements for effective function allocation (Feigh & Pritchett, 2014) and coactive design (Johnson, Bradshaw, Feltovich, et al., 2014) – showed that the outcomes of their design processes read like legal classifications of human-automated systems: what capacities are required to complete the task, how the human and automated agents can support each other's tasks, and what information

should be shared, with whom, and when. Framing classifications of human-automation systems through work allows for the classifications to stay relevant as capabilities evolve and change, and avoid the focus on specific amount of automation or conceptual location of the human which can limit innovation and create loopholes in the regulations.

3. Transparency: How do we determine and achieve the right levels of transparency for operators and regulators?

There has been an increasing call for transparency in automated and robotic systems. The reasons for this demand often come from concerns over trust, fault identification, and validation. A review of human-robot interaction and cognitive engineering showed that transparency has a deep and diverse literature, with many elements for further research but a significant foundation. To build a complete picture of transparency, we developed a synthesized model of six ways in which an automated system can be transparent: 1) how it evolves 2) how it learns, 3) its purpose, 4) its performance, 5) its software algorithms, and 6) its implementation on hardware. Relating transparency to trust revealed an inherent contradiction: for a human to trust human or non-human agents, there must be some opacity, some anthropomorphism of the Other to having internal states, not just hardware and lines of code. To trust is to be willing to be vulnerable whereas to distrust is to monitor and regulate. This trust is key to acceptance, and its rejection can increase workload and decrease safety. The types of transparency required for fault identification are fundamentally different than those required for trust. For automated systems to explain their faults or actions, the system will need to not only match content, but also the level of abstraction. The automated system must be able to recognize just when there is no sense in adding more detail to the explanation. Lastly, the transparency for validating systems that learn and evolve is an open-ended problem. There are increasing numbers of static and learning analyses for software, but there are many gaps in the capability to monitor and constrain learning, and a lack of clear guidelines from regulatory agencies.

4. Accountability: How should we determine responsibility for the actions of human-automation systems?

One of the main concerns with complex, human-automation systems is the possibility that after accidents, it will be too difficult to determine accountability for the failures and injuries. Narrowing this responsibility gap is not only an essential legal-ethical issue, but an important issue for cognitive engineers as well. We introduced methods for modeling and representing

accountability both at the early stage of design and after an accident. The function allocation design perspective attempts to explicitly model each agent's authorities for performing functions and their responsibilities for the outcomes (Pritchett et al., 2014a, 2014b). Mismatches occur when agents have responsibility for outcomes of actions but not the authority to perform the actions. The goal of the design process is to use mismatches to achieve desired redundancies but avoid mismatches that result in too much monitoring. For determining accountability after an accident, we described Rasmussen's (1997) risk management framework which has been used to understand several large-scale accidents. The framework focuses on balancing individual accountability with the system social and organizational factors. For a better understanding of the framework and its related methods, we then provided a case study of how the 2010 Flash Crash could be examined via the risk management framework. Given the capabilities of the post-accident methods like the risk management framework, with all their understandings of social and organizational factors, it is still an open question of how to translate these methods into the early stages of design to pre-emptively address accountability.

5. Safety: How do human-automated systems fail?

Safety concerns are one of the largest motivating factors for governing complex, sociotechnical systems, especially human-automated systems. We introduced the important safety principle of Defense-in-Depth, whose goal is to apply diverse safety barriers (defenses) along multiple points along accident scenarios to prevent single point failures (depth). These safety barriers can manifest themselves six general ways throughout a sociotechnical system: operational and maintenance, organizational and managerial, economic, research and education, regulator, and technical and design. Through the example of Turkish Airlines Flight 981 accident which caused the death of 346 people, we showed how each of the six sets of stakeholders can contribute to a single accident that seems like a straightforward design failure. The most common framework for mathematically describing safety and risk is quantitative risk assessment (QRA). While QRA has been successful in characterizing risk in many mechanical systems, its application is severely limited in human-automated systems. The most pervasive reason for this limitation is that human-automated systems are largely defined by their emergent behavior in which interaction between components of a system can cause outputs which cannot be predicted through only a deterministic understanding of the components. Human-automation interaction has innately emergent features which cause

significant challenges for modeling, simulation, and even, comprehension, which have yet to be fully addressed.

B. Conclusions

Considering the work in this paper, we have three main conclusions for our readers:

1. Be careful about assumptions regarding human-automated systems.

Within our research lab, the Cognitive Engineering Center, we have something called the "it-depends dance." The answer to almost every question about designing a human-automated system depends on countless aspects of the domain: what work is being done, by whom, why, how, within what constraints, norms, and ethics, etc.? The literature of cognitive engineering teaches us that given the nuances of human-automated systems, whether people should use common sense or not, well, depends. Sometimes common-sense does not apply. What may make sense from interactions with individual devices such as cell phones or computers, does not make sense for operating commercial aircraft with its coupled and complicated automated systems. Conversely, in other ways, common-sense is very useful in contradicting, particularly through analogy as shown in the discussions of trust and transparency in Sec. *IV. Transparency*.

2. Lawyers and ethicists should take advantage of the wealth of knowledge and experience within the science and engineering communities related to human-automated systems, like cognitive engineering.

The literature summarized in this paper are from diverse range of domains related to cognitive engineering (e.g., robotics, behavioral psychology, system safety, reliability engineering, and human factors), yet this is still only an introduction to the literature. There are many more frameworks, perspectives, definitions and theories that we did not include. Many of which ask the same questions as lawyers and ethicists, just from a technologist's perspective. Lawyers and ethicists can combine and synthesize methods at will, to form a valuable, grounded starting point without having to rely solely on their own experiences.

3. Effective governance of human-automated systems requires increased collaboration between those governing and those building these systems.

One reason for optimism regarding the recent legal-ethical frameworks is that most are formed from collaborations of lawyers, ethicists, policymakers, researchers, and technologists. We hope that those driving this legal-ethical renaissance related to human-automated systems continue to reach out to technologists. However, we stress that it is not just the lawyers, ethicists, and policymakers' responsibility to reach out to technologists. The fundamental goal of engineering and science is to better our society and without effective governance, the engineers and scientists will fail to achieve their goal (see (Vesilind, 2001)). It is also the obligation of those building these systems to engage the lawyers and policymakers on these issues. Therefore, we hope that our fellow engineers and scientists can, through this paper, see that their work is essential to legal, policy, ethical discussions of human-automated systems, and that there are still important questions with direct implications on the future of our society.

* * *

REFERENCES

Adams, D. (1997). The Hitchhiker's Guide to the Galaxy. *The Ultimate Hitchhiker's Guide*.

Alaieri, F., & Vellino, A. (2016). Ethical Decision Making in Robots: Autonomy, Trust and Responsibility Autonomy Trust and Responsibility. *International Conference on Social Robotics*. Springer International Publishing.

Bailey, R. W. (1982). *Human performance engineering: A guide for system designers*. Prentice Hall Professional Technical Reference.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779.

Bainbridge, L. (1997). The change in concepts needed to account for human behavior in complex dynamic tasks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *27*(3), 351–359. doi:10.1109/3468.568743

Baldwin, W. C., Felder, W. N., & Sauser. (2011). Taxonomy of increasingly complex Systems. *International Journal of Industrial and Systems*

Engineering, *9*(3), 298–316.

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, *23*(1), 4–17.

Belcastro, C. M. (2010). Validation and Verification of Future Integrated Safety- Critical Systems Operating under Off-Nominal Conditions. *AIAA Guidance, Navigation, and Control Conference*.

Belcastro, C. M. (2012). Validation and Verification (V&V) of Safety-Critical Systems Operating under Off-Nominal Conditions. *Optimization Based Clearance of Flight Control Laws* (pp. 399–419). Heidelberg: Springer Berlin.

Bennett, K. B., Posey, S. M., & Shattuck, L. G. (2008). Ecological Interface Design for Military Command and Control. *Journal of Cognitive Engineering and Decision Making*, *2*(4), 349–385. doi:10.1518/155534308x377829

Billings, C. E. (1997). *Aviation automation: The search for a human-centered approach*. (L. Erlbaum, Ed.). Mahway, NJ.

Bisantz, A. M., & Burns, C. M. (Eds.). (2009). *Applications of Cognitive Work Analysis*. CRC Press, Taylor & Francis Group.

Bisantz, A. M., Roth, E. M., Brickman, B., Gosbee, L. L., Hettinger, L., & McKinney, J. (2003). Integrating cognitive analyses in a large-scale system design process. *International Journal of Human-Computer Studies*, *58*(2), 177–206.

Bradshaw, J. M., Hoffman, R. R., Johnson, M., & Woods, D. D. (2013). The Seven Deadly Myths of "Autonomous Systems. *IEEE Intelligent Systems*, *13*, 2–9.

Breeman, G. (2012). Hermenuetic Methods in Trust Research. In F. Lyon, G. Mollering, & M. N. K. Saunders (Eds.), *Handbook of Methods on Trust* (pp. 149–160). Northampton, MA, USA: Edward Elgar.

Britcher, R. N. (1999). *The Limits of Software - People, Projects, and Perspectives*. Reading, MA: Addison Wesley Longman, Inc.

Buchanan, M. (2015). Trading at the speed of light. *Nature*, *518*(7538), 161.

Calo, R. (2014). A Horse of a Different Color: What robotics law can learn from cyberlaw. Slate. Retrieved from http://www.slate.com/articles/technology/future_tense/2014/10/robotics_law_should_take_cues_from_cyberlaw.html

Calo, R. (2016). *Robots in American Law* (No. 2016-04). University of Washington.

Canellas, M., & Haga, R. (2015). Toward Meaningful Human Control of Autonomous Weapons Systems through Function Allocation. *IEEE International Symposium on Technology and Society (ISTAS 2015)*. Dublin, Ireland: IEEE ISTAS.

Canellas, M., & Haga, R. (2016). Lost in translation: Building a common language for regulating autonomous weapons. *IEEE Technology and Society Magazine*, *35*(3), 50–58.

Cassano-Piche, A., Vicente, K., & Jamieson, G. (2009). A test of Rasmussen's risk management framework in the food safety domain: BSE in the UK. *Theoretical Issues in Ergonomics Science*, *10*(4), 283–304.

Castellano, G., Fanelli, A., & Torsello, M. (2007). Log data preparation for mining web usage patterns. *IADIS International Conference Applied Computing* (p. 20000).

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 15.

Chapter 482a - Autonomous Vehicles. (2013). State of Nevada.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency*. DTIC Document.

Chen, T., Kan, M., & Chen, X. (2015). TriRank : Review-aware Explainable Recommendation by Modeling Aspects. *CIKM 2015: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1661–1670. doi:10.1145/2806416.2806504

Courteney, H. (1999). Human factors of automation: the regulator's challenge. In S. Dekker & E. Hollnagel (Eds.), *Coping with computers in the cockpit*. Brookfield, VT: Ashgate.

Darling, K. (2017). Who"s Johnny?" Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. In P. Lin, G. Bekey, K. Abney, & R. Jenkins (Eds.), *ROBOT ETHICS 2.0*. Oxford University Press. doi:10.2139/ssrn.2588669

Darling, K., Nandy, P., & Breazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 770–775). doi:10.1109/ROMAN.2015.7333675

DARPA. (2016). Broad Agency Announcement: Explainable Artifical Intelligence (XAI. DARPA.

Dekker, S. W. A., & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human-automation co-ordination. *Cognition*.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Deployment of Autonomous Vehicles for Public Operation - California DMV. (2017). Retrieved from dmv.ca.gov

Desai, M., Stubbs, K., Steinfeld, A., & Yanco, H. a. (2009). Creating Trustworthy Robots: Lessons and Inspirations from Automated Systems. *Robotics Institute*. Retrieved from http://holman.cs.uml.edu/fileadmin/content/publications/2009/desai_paper.pdf

Directive 2007/46/EC of the European Parliment and of the Council. (2007). Official Journal of the European Union.

Docherty, B. L. (2012). *Losing Humanity: The Case Against Killer Robots*. Human Rights Watch.

Docherty, B. L. (2014). *Shaking the Foundations: The Human Rights Implications of Killer Robots*. Human Rights Watch.

Docherty, B. L. (2015). *Mind the Gap: The Lack of Accountability for Killer Robots*. Human Rights Watch.

Duijm, N. J. (2008). Safety-barrier diagrams as a safety management tool. *Reliability Engineering & System Safety*, *94*, 332–341.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human Computer Studies*, *58*(6), 697–718. doi:10.1016/S1071-5819(03)00038-7

Eddy, P., Potter, E., & Page, B. (1976). *Destination disaster: From the Trimotor to the DC-10*. New York: Time Books/Random House.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. doi:10.1037/0033-295X.114.4.864

FAA. (2016). Summary of Small Unmanned Aircraft Rule (Part 107). Federal Aviation Administration.

Federal Aviation Administration. (2013). *Operational Use of Flight Path Management Systems: Final Report of the Performance-based operations Aviation Rulemaking Committee/Commercial Aviation Safety Team Flight Deck Automation Working Group*.

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors*, *54*(6), 1008–1024.

Feigh, K. M., & Pritchett, A. R. (2014). Requirements for Effective Function Allocation A Critical Review. *Journal of Cognitive Engineering and Decision Making*, *8*(1), 23–32.

Felder, W. N., & Collopy, P. (2012). The elephant in the mist: What we don't know about the design, development, test and management of complex systems. *Journal of Aerospace Operations*, *1*, 317–327. doi:10.3233/AOP-120024

Feltovich, P. J., Bradshaw, J. M., Clancey, W. J., & Johnson, M. (2006). Toward an Ontology of Regulation: Socially-Based Support for Coordination in Human and Machine Joint Activity. *Pre-Proceedings of the Engineering Societies in the Agent's World 06 (ESAW06)*. Athens, Greece.

Fielder, J. H. (1992). Floors, doors, latches, and locks. In J. H. Fielder & D. Birsch (Eds.), *The DC-10 Case: A study in applied ethics technology*

and society. State University of New York Press.

Fielder, J. H., & Birsch, D. (Eds.). (1992). *The DC-10 Case: A study in applied ethics, technology, and society*. State University of New York Press.

Flight Deck Alerting System (FAS). (2007). SAE International.

Floyd, M. W., & Aha, D. W. (2016). Incorporating Transparency During Trust-Guided Behavior Adaptation. *International Conference on Case-Based Reasoning* (pp. 124–138). Springer.

Ford, C., & Jenks, C. (2016). The International Discussion Continues: 2016 CCW Experts Meeting on Lethal Autonomous Weapons. *Just Security*. Retrieved from https://www.justsecurity.org/30682/2016-ccw-experts-meeting-laws/

Foxall, G. R. (2005). Intentional behaviorism. *Understanding Consumer Choice* (pp. 173–198). Springer.

Future of Life Institute. (2017). *Asilomar AI Principles*. Retrieved from https://futureoflife.org/ai-principles/

Goillau, P., Kelly, C., Boardman, M., & Jeannot, E. (2003a). *Guidelines for Trust in Future ATM Systems : A Literature Review* (p. 70). EUROCONTROL. doi:HRS/HSP-005-GUI-02

Goillau, P., Kelly, C., Boardman, M., & Jeannot, E. (2003b). *Guidelines for Trust in Future ATM Systems : Measures* (p. 70). EUROCONTROL. doi:HRS/HSP-005-GUI-02

Greenblatt, N. A. (2016). Self-Driving Cars Will Be Ready Before Our Laws Are. *IEEE Spectrum*.

Grosz, B. J., Altman, R., Horvitz, E., Mackworth, A., Mitchell, T., Mulligan, D., & Shoham, Y. (2016). *Artificial Intelligence and Life in 2030*. One Hundred Year Study on Artificial Intelligence.

Haddon, D., & Whittaker, C. (2003). Aircraft airworthiness certification standards for civil UAVs. *The Aeronautical Journal (1968)*, *107*(1068), 79–86.

Hajdukiewicz, J. R., Burns, C. M., Vicente, K. J., & Eggleston, R. G. (1999). Work Domain Analysis for Intentional Systems. *Proceedings of the*

Human Factors and Ergonomics Society Annual Meeting, *43*(3), 333–337.

Hall, D. S., Shattuck, L. G., & Bennett, K. B. (2012). Evaluation of an Ecological Interface Design for Military Command and Control. *Journal of Cognitive Engineering and Decision Making*, *6*(2), 165–193.

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. doi:10.1177/0018720811417254

Harms-Ringdahl, L. (2009). Analysis of safety functions and barriers in accidents. *Safety Science*, *47*, 353–363.

Helldin, T. (2014). *Transparency for Future Semi-Automated Systems*. Orebro University.

Hettinger, A. Z., Roth, E. M., & Bisantz, A. M. (2017). Cognitive Engineering and Health Informatics: Applications and Intersections. *Journal of Biomedical Informatics*, 1–43.

Hill, R. W., Belanich, J., Lane, H. C., Core, M., Dixon, M., Forbell, E., Kim, J., et al. (2006). Pedagogically Structured Game-Based: Development of the Elect Bilat Simulation. *Army Science Conference*.

Holdren, J. P., & Smith, M. (2016). *Preparing for the Future of Artificial Intelligence*. Executive Office of the President, National Science and Technology Council, Committee on Technology.

Hollnagel, E. (2016). *Barriers and accident prevention*. Routledge.

Hollnagel, E., & Woods, D. D. (1983). Cognitive Systems Engineering: New Wine in New Bottles. *International Journal of Man-Machine Studies*, *18*, 583–600.

IATA. (2015). 2010-2014 Controlled Flight Into Terrain Accident Analysis Report. Montreal, Canada, and Geneva, Switzerland: International Air Transport Association.

IBM. (2017). Transparency and Trust in the Cognitive Era. Retrieved from https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-

principles/?ce=ISM0461&ct=stg&cmp=ibmsocial&cm=h&cr=storage&ccy=us

IEEE. (2016a). *Ethically Aligned Design: A Vision for Prioritizing Wellbeing with Artificial Intelligence and Autonomous Systems*. IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.

IEEE. (2016b). IEEE Announces Standards Development Project to Address Transparency of Autonomous Systems. Retrieved from http://standards.ieee.org/news/2016/ieee_p7001.html

Inagaki, T., Furukawa, H., & Itoh, M. (2005). Human Interaction with Adaptive Automation: Strategies for Trading of Control under Possibility of Over-Trust and Complacency. *HCI International 2005.Proceedings of the 11th International Conference on Human-Computer Interaction, Las Vegas, Nevada, USA, 2005*(September), 605–614. Retrieved from http://ezproxy.net.ucf.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ega&AN=ega213415&site=ehost-live

Jacklin, S. (2008). Closing the certification gaps in adaptive flight control software. *AIAA Guidance, Navigation and Control Conference and Exhibit* (p. 6988).

Jacklin, S., Schumann, J., Gupta, P., Richard, M., Guenther, K., & Soares, F. (2005). Development of advanced verification and validation procedures and tools for the certification of learning systems in aerospace applications. *Infotech@ Aerospace* (p. 6912).

Javaux, D. (2002). A method for predicting errors when interacting with finite state systems. How implicit learning shapes the user's knowledge of a system. *Reliability Engineering & System Safety*, *75*(2), 147–165. doi:http://dx.doi.org/10.1016/S0951-8320(01)00091-6

Jenkins, D. P., Salmon, P. M., Stanton, N. A., & Walker, G. H. (2010). A systemic approach to accident analysis: a case study of the Stockwell shooting. *Ergonomics*, *53*(1), 1–17.

Jenkins, D. P., Stanton, N. A., Salmon, P. M., & Walker, G. H. (2009). *Cognitive Work Analysis: Coping with Complexity*. Ashgate.

Jiancaro, T., Jamieson, G. A., & Mihailidis, A. (2013). Twenty Years of

Cognitive Work Analysis in Health Care: A Scoping Review. *Journal of Cognitive Engineering and Decision Making*, *8*(1), 3–22.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Riemsdijk, B. van, & Sierhuis, M. (2011). The fundamental principle of coactive design: Interdependence must shape autonomy. *Coordination, organizations, institutions, and norms in agent systems VI* (pp. 172–191). Springer.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Riemsdijk, M. B. van, & Sierhuis, M. (2014). Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of Human-Robot Interaction*, *3*(1), 43–69.

Johnson, M., Bradshaw, J. M., Hoffman, R. R., Feltovich, P. J., & Woods, D. D. (2014). Seven Cardinal Virtues of Human-Machine Teamwork: Examples from the DARPA Robotic Challenge. *IEEE Intelligent Systems*, *14*, 74–80.

Jones, M. L. (Ambrose). (2015). The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practies Principles. *Vanderbilt Journal of Entertainment and Technology Law*, *18*, 77–.

Kaplan, S. (1990). Bayes is for eagles. *IEEE Transactions on Reliability*, *39*, 130–131.

Kaplan, S., & Garrick, B. J. (1981). On The Quantitative Definition of Risk. *Risk Analysis*, *1*(1), 11–27. doi:10.1111/j.1539-6924.1981.tb01350.x

Kaszycki, M. (2014). Advisory Circular: Approval of Flight Guidance Systems (25.1329-1C). U.S. Department of Transportation, Federal Aviation Administration.

Kay, J., & Kummerfeld, B. (2012). Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(4), 24.

Kee, D., Jun, G. T., Waterson, P., & Haslam, R. (2017). A systemic analysis of South Korea Sewol ferry accident–Striking a balance between learning and accountability. *Applied Ergonomics*, *59*(504-516).

Klein, G. (2008). Naturalistic Decision Making. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 456–460.

Klyde, D. H., McRuer, D. T., & Myers, T. T. (1995). Unified Pilot-induced Oscillation Theory Volume I: Pio Analysis with Linear and Nonlinear Effective Vehicle Characteristics, Including Rate Limiting. Flight Dynamics Directorate, Wright Laboratory, Air Force Materiel Command.

Knuckey, S. (2014). Governments Conclude First (Ever) Debate on Autonomous Weapons: What Happened and What's Next. *Just Security*. Retrieved from https://www.justsecurity.org/10518/autonomous-weapons-intergovernmental-meeting/

Langley, D. Pat Meadows Ben Sridharan Mohan Choi. (2017). Explainable Agency for Intelligent Systems. *AAAI*. San Francisco.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function human-machine Systems. *Ergonomics*, *35*, 1243–1270.

Lee, J. D., & See, K. a. (2004). Trust in automation: designing for appropriate reliance. *Human factors*, *46*(1), 50–80. doi:10.1518/hfes.46.1.50.30392

Lee, K., Feron, E., & Pritchett, A. (2009). Describing Airspace Complexity: Airspace Response to Disturbances. *Journal of Guidance, Control, and Dynamics*, *32*(1), 210–222.

Levandowki, A. (2016). Statement from Anthony Levandowski on Self-Driving in San Francisco. Uber. Retrieved from https://newsroom.uber.com/statement-from-anthony-levandowski-on-self-driving-in-san-francisco/

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of management Review*, *23*(3), 438–458.

Lintern, G. (2012). Work-focused analysis and design. *Cognition, Technology & Work*.

Lomas, M., Chevalier, R., Cross II, E. V., Garrett, R. C., Hoare, J., & Kopack, M. (2012). Explaining robot actions. *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 187–188). ACM.

Mackay, W. E. (1999). Is paper safer? The role of paper flight strips in air

traffic control. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *6*(4), 311–340.

Madhavan, A. (2012). Exchange-traded funds, market structure, and the flash crash. *Financial Analysts Journal*, *68*(4), 20–35.

Marchant, G. E., Abbott, K. W., & Allenby, B. (2014). *Innovative Governance Models for Emerging Technologies*.

Marchant, G. E., Allenby, B. R., & Herkert, J. R. (2011). *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem* (Vol. 7). Springer Science & Business Media.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, *2*, 4–2.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709–734.

McIlroy, R. C., & Stanton, N. A. (2015). Ecological Interface Design Two Decades On: Whatever Happened to the SRK Taxonomy? *IEEE Transactions on Human-Machine Systems*, *45*(2), 145–163.

McRuer, D. T. (1995). Pilot-induced oscillations and human dynamic behavior. NASA.

Meier, M. W. (2016). U.S. Delegation Opening Statement (As delivered). *The Convention on Certain Conventional Weapons (CCW) Informal Meeting of Experts on Lethal Autonomous Weapons Systems*.

Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems: An introduction to computational models of social life*. Princeton: Princeton University Press.

Miller, M. J., & Feigh, K. M. (2017). Adressing the Envisioned World Problem: a case study in human spaceflight operations. *ACM Transactions on Computer-Human Interaction Submitted*.

Minotra, D., & Burns, C. M. (2016). Understanding safe performance in rapidly evolving systems: a risk management analysis of the 2010 US financial market Flash Crash with Rasmussen's risk management framework. *Theoretical Issues in Ergonomics Science*, 1–23.

Mirnig, A. G., Wintersberger, P., Sutter, C., & Ziegler, J. (2016). A framework for analyzing and calibrating trust in automated vehicles. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct* (pp. 33–38). ACM.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: vigilance and task complexity effects. *Human Factors*, *38*, 311–322.

Muethel, M., & Hoegl, M. (2012). The influence of social institutions on managers' concept of trust: Implications for trust-building in Sino-German relationships. *Journal of World Business*, *47*(3), 420–434.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man-Machine Studies*, *27*, 527–539. doi:10.1016/S0020-7373(87)80013-5

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460.

Müller, V. C. (2016). Drones and Responsibility: Legal, Philosophical, and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons. In E. D. Nucci & F. S. de Sio (Eds.), (pp. 1–16). London: Ashgate.

Murphy, R. R., & Shields, J. (2012). *Task Force Report: The Role of Autonomy in DoD Systems*. Department of Defense: Defense Science Board.

Newnam, S., & Goode, N. (2015). Do not blame the driver: a systems analysis of the causes of road freight crashes. *Accident Analysis & Prevention*, *76*, 141–151.

NHTSA. (2016). *Federal Automated Vehicles Policy - U.S. Department of Transportation, National Highway Transportation Safety Administration*.

NHTSA, U. S. (1999). Federal Motor Vehicle Safety Standards and Regulations. US Department of Transportation.

Norman, D. A. (1990). The 'problem' with automation: inappropriate feedback and interaction, not'over-automation. *Philosophical*

Transactions of the Royal Society B: Biological Sciences, *327*(1241), 585–593.

NSTC. (2016). The National Artificial Intelligence Research and Development Strategic Plan. National Science and Technology Council.

Orasanu, J., & Connolly, T. (1993). The Reinvention of Decision Making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision Making in Action: Models and Methods* (pp. 3–20). Norwood, NJ: Ablex Publishing Corporation.

Owotoki, P., & Mayer-Lindenberg, F. (2007). Transparency of Computational Intelligence Models. *Research and Development in Intelligent Systems XXIII* (pp. 387–392). Springer.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381–410.

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.

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. doi:10.1109/3468.844354

Partnership on AI. (2017). *Tenants: Partnership on AI to benefit people and society*. Retrieved from https://www.partnershiponai.org/tenets/

Poggio, T. (2012). The levels of understanding framework, revised. *Perception*, *41*(9), 1017–1023.

Pritchett, A. R. (2009). Aviation Automation: General Perspectives and Specific Guidance for the Design of Modes and Alerts. *Reviews of Human Factors and Ergonomics*, *5*(1), 82–113. doi:10.1518/155723409X448026

Pritchett, A. R., Kim, S. Y., & Feigh, K. M. (2014a). Measuring Human-Automation Function Allocation. *Journal of Cognitive Engineering and Decision Making*, *8*, 52–77.

Pritchett, A. R., Kim, S. Y., & Feigh, K. M. (2014b). Modeling Human–Automation Function Allocation. *Journal of Cognitive Engineering and Decision Making*, *8*(1), 33–51.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety science*, *27*(2), 183–213.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. New York: Wiley.

Rausand, M., & Hoyland, A. (2003). *System Reliability Theory: Statistical Methods and Applications*. Wiley.

Read, G. J. M., Salmon, P. M., & Lenne, M. G. (2015). Cognitive work analysis and design: current practice and future practitioner requirements. *Theoretical Issues in Ergonomics Science*, *16*, 154–173.

Reason, J. (1997). *Managing the Risks of Organizational Accidents*. Brookfield, VT, USA: Ashgate.

Richards, N. M., & Smart. (2013). How Should the Law Think About Robots? SSRN. Retrieved from https://ssrn.com/abstract=2263363

Riley, V. (1996). What avionics engineers should know about pilots and automation. *Aerospace and Electronic Systems Magazine, IEEE*, *11*(5), 3–8.

Riley, V. (2000). Developing a Pilot-Centered Autoflight Interface. *2000 World Aviation Conference*. Reston, Virigina: American Institute of Aeronautics and Astronautics.

Riley, V., DeMers, B., Misiak, C., & Schmalz, B. (1999). A pilot-centered autoflight system concept. *Aerospace and Electronic Systems Magazine, IEEE*, *14*(9), 3–6.

Riley, V., DeMers, B., Misiak, C., & Shackleton, H. (2002). The Cockpit Control Language Program: An Update. *SAE International*, *111*, 561–566.

SAE International. (2016). *J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*.

Saleh, J. H., Marais, K. B., Bakolas, E., & Cowlagi, R. V. (2010). Highlights

from the literature on accident causation and system safety: Review of major ideas, recent contributions, and challenges. *Reliability Engineering & System Safety*, *95*(11), 1105–1116. doi:http://dx.doi.org/10.1016/j.ress.2010.07.004

Saleh, J. H., & Pendley, C. C. (2012). From learning from accidents to teaching about accident causation and prevention: Multidisciplinary education and safety literacy for all engineering students. *Reliability Engineering & System Safety*, *99*, 105–113. doi:http://dx.doi.org/10.1016/j.ress.2011.10.016

Selkowitz, A., Lakhmani, S., Chen, J. Y. C., & Boyce, M. (2015). The Effects of Agent Transparency on Human Interaction with an Autonomous Robotic Agent. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 806–810. doi:10.1177/1541931215591246

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of human factors and ergonomics*, *1*(1), 89–129.

Sheridan, T. B., & Verplank, W. L. (1978). Human and Computer Control of Undersea Teleoperators. MIT.

Sklet, S. (2006). Safety barriers: Definition, classification, and performance. *Journal of Loss Prevention in the Process Industries*, *19*(5), 494–506. doi:http://dx.doi.org/10.1016/j.jlp.2005.12.004

Smith, B. W. (2016). Uber vs. The Law. The Center for Internet and Society. Retrieved from http://cyberlaw.stanford.edu/blog/2016/12/uber-vs-law

Smith, B. W., Svensson, J., Humanes, P., Konzett, G., Paier, A., Mizuno, T., & Lykotrafiti, A. (2015). Automated and Autonomous Driving: Regulation under uncertainty. Organization for Economic Co-operation and Development.

Sorensen, J. N., Apostolakis, G. E., Kress, T. S., & Powers, D. A. (1999). On the role of defense in depth in risk-informed regulation. *Proceedings of PSA '99, International Topical Meeting on Probabilistic Safety Assessment*, *99*, 22–26.

Spirkovska, L., Iverson, D., Hall, D., Taylor, W., Patterson-Hine, A., Brown, B., Ferrell, B., et al. (2010). Anomaly Detection for Next-Generation

Space Launch Ground Operations. *SpaceOps 2010 Conference Delivering on the Dream Hosted by NASA Marshall Space Flight Center and Organized by AIAA* (p. 2182).

Testing of Autonomous Vehicles - State of California, Department of Motor Vehicles. (2017). Retrieved from https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/testing

The FAA and Industry Guide to Product Certification. (2004). FAA.

UN CCW. (2016). *Reccomendations to the 2016 Review Conference on Lethal Autonomous Weapons (Submitted by the Chairperson of the Informal Meeting of Experts, United Nations Comittee on Certain Conventional Weapons)*.

Vaina, L. M., & Passingham, R. E. (2016). *Computational Theories and their Implementation in the Brain: The legacy of David Marr*. Oxford University Press.

Vesilind, P. A. (2001). Engineering as Applied Social Science. *Journal of Professional Issues in Engineering Education and Practice*, *127*(4), 184–188.

Vicente, K. J. (1999). *Cognitive Work Analysis: Toward safe, productive & healthy computer-based work*. Lawrence Erlbaum Associates.

Vicente, K. J., & Christoffersen, K. (2006). The Walkerton E. coli outbreak: a test of Rasmussen's framework for risk management in a dynamic society. *Theoretical Issues in Ergonomics Science*, *7*(02), 93–112.

Walker, G. H., Stanton, N. A., Salmon, P. M., & Jenkins, D. P. (2008). A review of sociotechnical systems theory: a classic concept for new command and control paradigms. *Theoretical Issues in Ergonomics Science*, *9*(6), 479–499.

Wang, Q. (2012). *Developing a Computational Framework for Explanation Generation in Knowledge-based Systems and its Application in Automated Feature* (No. May). RMIT.

Waterson, P., Robertson, M. M., Cooke, N. J., Militello, L. G., Roth, E. M., & Stanton, N. A. (2015). Defining the methodological challenges and opportunities for an effective science of sociotechnical systems and safety. *Ergonomics*, *58*(4), 565–599.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117.

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics*, *23*, 995–1011.

Witte, J. B. (2004, March). *An investigation relating longitudinal pilot-induced oscillation tendency rating to describing function predictions for rate-limited actuators*. Air Force Institute of Technology.

Woods, D. D. (1985). Cognitive technologies: The design of joint human-machine cognitive systems. *AI magazine*, *6*(4), 86.

Woods, D. D., & Dekker, S. W. A. (2000). Anticipating the effects of technological change: a new era of dynamics for human factors. *Theoretical Issues in Ergonomics Science*, *1*(3), 272–282.

Woods, D. D., & Hollnagel, E. (2006a). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. Pattersn in Cognitive Systems Engineering. Boca Raton, FL: Taylor & Francis.

Woods, D. D., & Hollnagel, E. (2006b). *Joint Cognitive Systems*. Pattersn in Cognitive Systems Engineering. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Woods, D. D., & Roth, E. M. (1988). Cognitive engineering: Human problem solving with tools. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *30*(4), 415–430.

Zhu, J. (2009). Intentional systems and the artificial intelligence (AI) hermeneutic network: Agency and intentionality in expressive computational systems. *ProQuest Dissertations and Theses*, (August), 251.