Matters of Ethics, Trust, and Potential Liability for Autonomous Systems

J. Christopher Brill[1], James P. Bliss[1], Peter A. Hancock[2], Dietrich Manzey[3], Joachim Meyer[4], & Alison Vredenburgh[5]

[1] *Old Dominion University,* [2] *University of Central Florida,* [3] *Berlin Institute of Technology*
[4] *Tel Aviv University,* [5] *Vredenburgh & Associates, Inc.*

Abstract – The objective of this panel was to discuss issues related to the development and use of autonomous systems, with specific focus on the overriding themes of ethical considerations and potential liability for Human Factors and Ergonomics (HF/E) professionals who are involved in their development. *Chris Brill* provided opening remarks to frame the discussion and introduce the panelists. *James Bliss* discussed legal implications related to our collective penchant for developing conservative, false-alarm prone automation. *Peter Hancock* advocated for human-centered constraints on autonomous systems, as they may, one day, pose an existential threat to humanity. *Dietrich Manzey* discussed ethical considerations for autonomous systems, including how design can encourage ethical user behavior. *Joachim Meyer* argued that HF/E professionals have an obligation to help designers understand the ethical implications of poor design, particularly in the context of autonomous systems. Lastly, *Alison Vredenburgh* provided thoughts on potential liability for HF/E professionals, particularly in light of the relative newness of autonomous systems. The panel then turned to facilitated discussion with panelists and audience members. Specific themes included the boundaries of our responsibilities as HF/E professionals for ill-conceived or morally-objectionable systems, potential implications of manipulating user trust through design, cross-cultural perspectives on public acceptance and legal peril, and how concerns might differ by domain (e.g., medical vs. combat vs. manufacturing). The session concluded with panelists summarizing how ethics influence design and recommendations for how HF/E professionals can potentially protect themselves from legal liability for mishaps involving autonomous systems they helped develop.

INTRODUCTION

A Nearly-Blind Sprint into the Brave World of Autonomous Systems

J. Christopher Brill, Panel Organizer
Old Dominion University

This inspiration for this discussion panel proposal stemmed from recent articles on autonomous systems dealing with potential concerns regarding safety, ethics, and legal liability. Numerous articles have been published on autonomous driving systems, one of which details a team's 2994 mile drive from Los Angeles to New York City in less than 59 hours (including charging time) using Tesla's semi-autonomous "autopilot" feature (Davies, 2015). Automation was engaged 96% of the time and included segments driven at speeds approaching 90 miles per hour. The author questions whether users should even be capable of activating autopilot at such high speeds. As HF/E professionals, we advocate for the user experience and system flexibility, but we also understand the value and necessity of constraining actions - particularly when safety is concerned. Indeed, system designers do their employers a disservice by not reining in reasonably foreseeable unsafe user behavior, thereby, leaving companies open to litigation.

Conversely, Google's experimental self-driving cars may have the opposite problem - too many constraints on the user, as the most recent design lacks user-accessible controls (i.e., steering wheel, brake and acceleration pedals; Prynn, 2014). Proponents of the system claim this omission eliminates human users as a dangerous source of unpredictability in the system (Lavrinc, 2014). By design, users *cannot* assume manual control in the event of automation failures. Some may find this disturbing; during a 14-month test period, Google's engineers had to reclaim manual control of experimental test vehicles (in versions of the car that had manual controls) 341 times (Davies, 2014). Although most instances were due to mechanical failure, 20% occurred due to poor judgment or decision-making on the part of the automation. Notwithstanding, even if an autonomous vehicle has manual controls, passengers may be poorly poised to assume control. Their attention may be otherwise occupied by smart phone usage, or they may have been lulled into an altered conscious state by the car's motion through a natural vestibular response, called sopite syndrome (Graybiel & Knepton, 1976; Lawson & Mead, 1998). Indeed, the very role of "passive passenger" makes sopite syndrome's assertion all the more likely, leading to sleepiness despite receiving and adequate night's rest, disinclination to stay on task, fuzzy-headedness, and general malaise - all of which could contribute to increased automation complacency and reliance.

Another article questions whether developers of autonomous weapons systems can be prosecuted for war crimes (Majumdar, 2014). The issue of autonomous weapons is an ethical maelstrom, with proponents and vocal opponents. The question of war crime prosecution is an interesting one, and it's a very safe bet that the first major mishap involving unintended civilian deaths by an autonomous weapons system will result in *extensive* inquiry to determine its cause. HF/E professionals who were part of the development team will need to account for their roles and inputs (or lack thereof) into its design, as the system and its full history will be stripped

back and examined, screw-by-screw, from proverbial smoking barrel to system inception. Although the end user (e.g., operator or commander on battleground) is at the sharp end of the stick, latent factors will likewise receive scrutiny. A wise HF/E professional will need to surefooted on his or her role in system development. For this type of situation, documentation of scientifically-sound design recommendations, data, and written evidence of any safety or performance concerns will likely provide the best defense.

Autonomous medical systems (virtual doctors), farming systems, commercial shipping, and legal services will also be available in the near future (Al-Khatib, 2016), demonstrating near-term ubiquity. Notwithstanding, there remains a significant, ongoing need for HF/E professionals. We must be open to their benefits and apply scientific principles to facilitate the best possible performance, while engendering an appropriate level of user trust. Likewise, we must sound the alarm when system trust is disproportionate or unwarranted. Perhaps most importantly, we must also serve as a conscience for the technology development community, for we understand the perils of poorly implemented or ill-considered design. Dr. Ian Malcolm, of the movie *Jurassic Park*, may have captured this sentiment best: "... your scientists were so preoccupied with whether or not they *could* that they didn't stop to think whether they *should*." If we decide we should (develop a particular autonomous system), then it is our responsibility to employ our scientific prowess to mold and nurture it, and when necessary, constrain or oppose it.

J. Christopher Brill, Ph.D., is a Senior Research Psychologist and Team Lead for the Human Insight and Trust Team at the Air Force Research Laboratory. He previously held faculty positions at Old Dominion University and Michigan Tech. His research focuses on human-system trust, multimodal displays, and human performance assessment. The present work was submitted for publication prior to his employment with the U.S. government; the views and opinions contained herein are those of the author and should not be construed as an official U.S. government position, policy, or decision.

Automation and Products Liability: Implications for System Trust and Litigation

James P. Bliss
Old Dominion University

Researchers and task designers have become very interested in the impact of increasing automation on trust. Clearly, reduced levels of operator trust can degrade ongoing and future task performance, as demonstrated by many researchers (Breznitz, 1984; Getty, Swets, Pickett, & Gonthier, 1985). Performance effects have included slowed reaction time and impaired accuracy. Though such results may occur partly because of loss of situation awareness or changes in cognitive workload, many researchers have isolated trust as a vulnerable construct, especially when automation is clumsy (Weiner, 1989).

Legal decisions have for many years impacted the use of automated sensor based warning systems. Embracing a "better safe than sorry" approach, designers have often approached the creation and implementation of automated systems from the perspective of a duty to warn. As a result, many modern systems present users with excessive warning signals, even when the statistical chance of a true problem is low. One salient example is the low tire pressure warning system in automobiles. Common criticisms of the system include false alarms due to outside temperature (Crowell, 2015), damage from tire-mounting machines (Allen, 2009), or sensor batteries that have failed (Carley, 2015).

As the domain of environments for automated systems becomes more diversified, the frequency and consequences of automation failure will rise. Because automated systems are designed, fielded products, the potential for personal injury, and associated legal challenge, exists. As human factors researchers make recommendations to improve the design and implementation of automated systems, careful consideration of the legal ramifications of automation failure is warranted. This presentation will focus on three of the significant concepts of products liability law: strict liability, negligence, and breach of warranty. Each will be discussed as they pertain to modern automated systems.

James P. Bliss, Ph.D., joined the Psychology Department at Old Dominion University as an Associate Professor in 2001. He was awarded tenure in 2004, and is currently the department chair. He and his students conduct research in two primary areas. One is the occurrence of alarm (and automation) mistrust. Specifically, he is interested in what factors contribute to the development of mistrust, and how designers and trainers can optimize compliance to automated systems. The second broad area includes the use of virtual environments for task training. Dr. Bliss and his students have worked with a number of research institutions, including Boeing, DARPA, the Office of Scientific Development, the US Army, and NASA, and AFOSR. Dr. Bliss has also served as an expert witness in a number of product liability cases involving warnings and alarms.

Designing Limits to Autonomous Systems

Peter A. Hancock
University of Central Florida

We are witnessing an approaching sea-change in the way that advanced technological systems operate. Embedded in this line of evolution of our current automation there is a growing wave of ever-more independent, autonomous systems. The degree of interaction between such autonomy with any human operator is becoming progressively diminished. With this decreasing interaction comes decreasing system control on behalf of any human agency. In this presentation, I will advocate for human-centered constraints to be designed, programmed, promulgated and imposed upon these nascent forms of independent entity. It is important in the beginning to define the relevant terms. Here, I define automation as: '*those systems designed to accomplish a specific set of largely deterministic steps, most often in a repeating pattern, in order to achieve one of an envisaged and limited set of pre-defined outcomes*.' In apparent contrast, and I emphasize the word apparent here, I define autonomous

systems as: '*those systems which are generative and learn, evolve and permanently change their functional capacities as a result of the input of operational and contextual information. Their actions necessarily become more indeterminate across time*' (see Hancock, 2016). What is very clear from the vector of our present progress is that that nascent growth of autonomy is necessarily predicated upon increasing levels of automation. In consequence, the two terms are certainly not in any form of contrast or mutual contradiction. Rather, they represent differing serial stages of computational evolution. Like the evolutionary rates of all species, some facets of automation will rapidly change into autonomy. In other contexts simple automation will continue to suffice for the required task at hand. In these later cases, little impetus will drive such a specific system to any higher level of complexity and/or autonomy. The landscape of our own natural world provides a facile analog of this overall mimetic, inter-species existence. As the only fundamental difference between the conceptions of Lamarck and Darwin lie in the passage of time, the fact that technological evolution most closely approximates Lamarck's vision does not mean species evolutionary principles are inapplicable in this latter case of these human fabricated technological systems. They simply change more quickly and, putatively, in more purposive and logical sequences. Since their actions necessarily become more opaque across time as their computational capacities increase we must ask crucial questions now about trust, reliability, and determinacy of action. This requires that we set design boundaries and constraints as these technologies evolve. Since their rate of evolution will outstrip our own by many orders of magnitude, the time to enact those design constraints is now. The fact that such design constraints will not be is simply another expression of the vacuity of the human species; a vacuity evident in its everyday individual and collective actions.

Ergonomics and Human Factors are disciplines which are purpose-directed to mediate between human beings and the machines they create. For most of their existence, tools and their later expression in more advanced technologies have served to render human visions into material form in order, nominally, to facilitate individual and collective well-being. That the genesis of such technology and much of its associated progress has derived from the conflictive nature of the human character remains problematic. Humans and their propensity toward individual and small group optimization are often in conflict with the goals and aspirations of the putative 'other' (broadly defined) and this has proved to be no great cause for celebration. The wars with these others, whether our distant kin, other living things, and now the very environment that sustains us have seen great expansions of such material visions. Yet today many of these advances stand in direct opposition to, and threaten the continued existence of, our whole species. Now the very technological vehicles of 'progress' are providing fundamental, profound and even existential threats. Whether technical solutions are necessarily required for perceived technical problems represents the quandary of our times. Here, I vacillate between the cheap, tawdry aspirations of delusional optimism and the vast, dark reality of rational pessimism. Please come and experience my personal, classical clasticism.

Peter A. Hancock, D.Sc., Ph.D. is *Provost Distinguished Research Professor* in the Department of Psychology and the Institute for Simulation and Training, as well as at the Department of Civil and Environmental Engineering and the Department of Industrial Engineering and Management Systems at the University of Central Florida (UCF). He directs the MIT^2 Research Laboratories and Associate Director of the Center for Applied Human Factors in Aviation (CAHFA). Professor Hancock is the author of over seven hundred refereed scientific articles and publications, as well as writing and editing fifteen books. He has been continuously funded by extramural sources for every one of the thirty-one years of his professional career. To date, he has secured over $17 Million in externally-funded research during his career.

Ethics, accountability for (mis-)use of automated systems and the issue of automation surprise

Dietrich Manzey
Berlin Institute of Technology

In my contribution, I will address three aspects which I consider important issues with respect to ethics of design of automated and/or autonomous systems.

First, what is the ethical basis for decisions to become involved in some technology development, and what does that mean for the individual accountability? This sort of ethical issue involves ethical considerations and decisions on a societal as well as individual level. In democratic societies, the first level usually involves what I would term informed (political) discussions about general aspects of technology use. A recent example is the decision in Germany to step out of the nuclear power program, based on considerations about the limits of controllability and risks involved in this sort of technology. Typically such discussions results guidelines (laws, regulations) about what technology is acceptable and what not. However, does this release from any ethical considerations and accountability of becoming involved in technology development which is considered as acceptable? No, certainly not. The mere fact that a society accepts certain technologies and systems does not release an individual designer or manufacturer to decide about becoming involved. This always involves personal considerations of ethics and morals, which then lead to an individual decision to become involved or not. In case of a positive decision, this also includes to take the responsibility and accountability for the consequences of use of the technology. For example, if you decide to become involved as human factors engineer in the design of technology or interfaces of systems that might harm persons, you inevitably also share accountability for the use of these systems and the consequences. This holds true although you might have been involved only as a scientist. The paradigmatic historical example is that of the (German) scientist involved in the atomic bomb program first in Germany and later in the US.

Secondly, as we know from our human factors research, there are several issues in human-automation interaction that

relate to incidents of overtrust or, more general, misuse of automation. Most discussed examples involve complacency and automation bias (Parasuraman & Manzey, 2010). What does that mean for the ethics of automation design and the accountability of engineers and human factors experts involved in the design of these systems? I will argue that all of our concepts of human-centered system design might provide a sort of ethical guideline for most of these issues. However, there might also emerge specific ethical issues related to the use of automated systems which go beyond this and which need to be carefully to be considered by designers in addition to the basic principles of human-centered design. For example, Cummings (2006) has discussed the issue of interfaces to serve as moral buffer for human actions which particularly represent an (ethical) issue when developing interfaces of systems to be used in the military or medical domain that can harm people.

Thirdly, complex automated systems like currently used in aviation but also other domain has repeatedly shown to provide what has been referred to as *automation surprises* (Sarter et al., 1997). Often these "surprises" result from situations and complex constellations which, in principle, do not seem to be anticipatable or avoidable by neither the manufacturer of the system, nor the user. They directly correspond to what has been referred to as normal accidents by Perrow (1999). From my point of view, this raises a complex ethical issue of responsibility and accountability which is difficult to resolve. Take, for example, an accident occurring with a state-of-the art airplane manufactured according to the highest engineering and human factors standard, operated by highly trained pilots of a major airline with highest safety standards. Who is accountable or liable for the accident? The pilot? The manufacturer? The airline? Any others? I will argue that, similar to the discussion around normal accidents, any questions of individual accountability or liability are misaddressed in this case. Ethically, we probably have to live with such instances and their consequences if we, as society or humanity, decide to introduce and use this sort of technology. If at all, the society might be taken accountable for the consequences resulting from this decision.

Dietrich Manzey, Ph.D,. is a university professor of work, engineering, and organizational psychology at Technical University of Berlin (TU Berlin), Germany. He received his Ph.D. in experimental psychology at the University of Kiel, Germany, in 1988. Among others, his research interests include issues of human-automation interaction, system safety in high-hazard industries, and multitasking and human performance in extreme environments.

Doing Good with Good Human Factors

Joachim Meyer
Tel Aviv University

The design of the user side of technologies, the focus on users' interactions with systems, and the need to define users' role in advanced systems inherently touch on ethical issues. Some of these issues are classic ethical dilemmas. Do we design the system so that the overall utility from using the system will be maximal (as prescribed by an utilitarian position, proposed by Bentham, Mill, and more recently, Harsany) or should we design the system so that the weakest, least advantaged user can benefit from it (a Rawlsian approach)? When we design a system, should we focus on the users' interests or should we prefer the interests of the organization that hired us? Should we, as a profession, define the limits of ethically sound applications of human factors, or should we leave it to the individual practitioners to decide what is acceptable for them?

Here, as in any other case when one faces ethical dilemmas, it is impossible to provide definite, universally accepted guidelines. However, I argue that it is the duty of the human factors professional to contribute to the discussions of ethical points. The application of sound human factors knowledge can be valuable input and can often serve as a "sanity check" for decisions regarding ethical issues.

To provide the needed input for the discussions, the profession must have tools to provide clear predictions of human behavior and system performance, given certain decisions. We should be able to predict that if one decides to automate function X, there will be a Y% chance of a malfunction, which can lead to some negative consequence (*N* people will be hurt). If one chooses not to automate the function, there will be some probability for a malfunction too, and either more or fewer people will be hurt. The decision whether these are acceptable values, given the costs and investments needed for the different alternatives, will probably be made by others. However, we should be the ones to provide the input for these decisions.

At times we can rule out certain decisions. For instance, if an operator is expected to monitor a system passively for hundreds of hours in order to intervene within seconds when a malfunction occurs, we can safely say that this operator is doomed to fail. Such a statement may not be aligned with the preferences of the legal advisors of the organizations. They may prefer to blame the specific operator for problems of the system, rather than the system owner or designer. However, our knowledge and expertise can help prevent such bad designs. If such a design is still implemented (perhaps because no human factors professional was involved in the design process), we should be able to inform whoever investigates incidents what was and was not likely to happen, given the situation and the characteristics of the system and the user. This is already widely done in the context of traffic safety and forensic human factors, but we still haven't gotten very far in the context of automation.

Thus, I would argue that the role of human factors professionals and cognitive engineers is to present the implications of design and operations decisions. The decisions will eventually then be made by others, but ideally the input should have some importance. To face up to this task, we have to generate sound knowledge and validated models to predict the consequences of design and operations decisions. Without them our recommendations will carry little weight, and we may even at times cause damage.

Joachim Meyer, Ph.D., is a professor in the Department of Industrial Engineering at Tel Aviv University. He is the

current department chair and the founder of the Interacting with Technology (IwiT) laboratory. During the 2014-15 academic year, he was on sabbatical in Boston, where he served as a visiting professor with the Human Dynamics Group at the MIT MediaLab. He holds an M.A. in Psychology and a Ph.D. in Industrial Engineering from Ben-Gurion University. He is an associate editor for IEEE Transactions on Human Machine Systems and for the Journal of Cognitive Engineering and Decision Making and is on the editorial board of Human Factors. His primary areas of research include decision aids and warning systems, interaction with adaptive automation, human-robotic interaction, and medical decision-making.

Forensics Considerations Regarding Autonomous Systems
Alison Vredenburgh
Vredenburgh & Associates, Inc.

Human factors professionals who provide a forensic analysis of products consider how manufacturers manage hazards. This evaluation is through the lens of the hazard management and control hierarchy that addresses risk management through design, barriers/guarding, and warnings and risk communication systems. When analyzing a product, an important consideration is its foreseeable uses and misuses. This means that products need to be evaluated by testing the range of potential user populations in the expected use environments. Part of determining potential misuses is to consider transfer of training from older technologies to autonomous systems; a negative transfer would occur when new systems require user responses inconsistent with prior learned behavior (Vredenburgh & Zackowitz, 2006).

User expectations are a key factor in risk perception and thus whether users read and comply with safety information. Warning effectiveness tends to increase as a function of perceived hazardless; if autonomous products appear to be safer, users may take fewer self-protective measures (Vredenburgh & Zackowitz, 2006). Therefore, strongly worded warnings may be needed to override these perceptions. Moreover, it is important to consider that marketing materials and advertising for these new technologies may act as anti-warnings; media that represent dangerous products as safer can contradict or undermine warnings (Bohme & Egilman, 2006). Anti-warnings may make new technologies appear to eliminate or reduce the need for monitoring and safe behavior by the users.

Autonomous systems are a reasonably new technology; thus expectations based on prior experience with older technologies of similar products and safety perceptions about this new market segment would be key factors for consideration in a forensic evaluation.

Alison Vredenburgh, Ph.D., is the principal of Vredenburgh & Associates, Inc. She has published more than eighty peer-reviewed papers in the areas of human factors, safety, and psychology. She has served as an expert witness for hundreds of personal injury and products liability cases and has testified in Municipal, Superior, and Federal District Courts. She has held several offices within HFES and is the current Program Chair of the Forensics Professional Group. She holds a Ph.D. in Industrial-Organizational Psychology and is a Certified Professional Ergonomist (CPE).

REFERENCES

Al-Khatib, T. (2016). *Robo-lawyers, farmers, docs: our future on autopilot.* http://news.discovery.com/tech/robotics/robo-lawyers-farmers-docs-our-future-on-autopilot-160204.htm. Retrieved February 6, 2016.

Allen, M. (2009). *How to troubleshoot a tire-pressure monitoring system.* Retrieved from http://www.popularmechanics.com/cars/how-to/a4849/4336449/. January 7, 2016.

Bohme, S.R., & Egilman, D. (2006).Consider the source: Warnings and anti-warnings in the tobacco, automobile, beryllium and pharmaceutical industries. In M. Wogalter (Ed.), *The Handbook of Warnings*. (pp. 635-667). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Breznitz, S. (1984): Cry wolf: *The psychology of false alarms*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carley, L. (2015). *Diagnostic strategies for solving TPMS errors.* Retrieved from http://www.brakeandfrontend.com/tpms-diagnostic-strategies-for-solving-errors/. January 7, 2016.

Crowell, C. (2015). *Winter cold wreaks havoc on TPMS sensors.* Retrieved from http://www.tirereview.com/tpms-sensors-cold-weather/. January 7, 2016.

Cummings, M.L. (2006). Automation and accountability in decision support system interface design. *Journal of Technology Studies*, *32*, 23-31.

Davies, A. (2015). *Obviously drivers are already abusing Tesla's autopilot.* Retrieved from http://www.wired.com/2015/10/obviously-drivers-are-already-abusing-teslas-autopilot/. January 7, 2016.

Graybiel, A., & Knepton, J. (1976). Sopite syndrome: a sometimes sole manifestation of motion sickness. *Aviation, Space, and Environmental Medicine*, *47*(8), 873-882.

Getty, D.J., Swets, J.A., Pickett, R.M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1*(1), 19-33.

Hancock, P.A. (2016). *Imposing limits on autonomous systems*. Manuscript in preparation.

Lavrinc, D. (2014). *The real reason Google's self-driving car doesn't have controls.* Retrieved from http://jalopnik.com/the-real-reason-googles-self-driving-car-doesnt-have-co-1583056841. January 4, 2016.

Lawson, B. D., & Mead, A. M. (1998). The sopite syndrome revisited: drowsiness and mood changes during real or apparent motion. *Acta astronautica*, *43*(3), 181-192

Majumdar, D. (2014). *Essay: The legal and moral problems of autonomous strike aircraft.* Retrieved from http://news.usni.org/2014/08/21/essay-legal-moral-problems-autonomous-strike-aircraft. January, 3, 2016.

Parasuraman, R. & Manzey, D. (2010). Complacency and bias in human use of automation: A review and attentional synthesis. *Human Factors, 52*, 381-410.

Perrow, C. (1999). *Normal accidents. Living with high risk technology.* Princeton: Princeton University Press.

Prynn, J. (2014). *Google unveils 'self driving' car with no controls other than start button.* Retrieved from http://www.standard.co.uk/news/techandgadgets/google-unveils-self-driving-cars-with-no-controls-other-than-start-button-9443732.html. January 7, 2016.

Sarter, N. B., Woods, D. D., and Billings, C. E. (1997). Automation surprises. In Salvendy, G. (ed.), *Handbook of human factors and ergonomics.* 2nd ed. John Wiley and Sons.

Vredenburgh, A.G., & Zackowitz, I.B. (2006). Expectations. In M. Wogalter (Ed.), *The Handbook of Warnings*. (pp. 345-354). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Weiner, E.L. (1989). *Human factors of advanced technology ("glass cockpit") transport aircraft*. (NASA Contractor Report No. 177528). NASA Ames Research Center.