Building a framework to manage trust in automation

J.S. Metcalfe*[a], A.R. Marathe[a], B. Haynes[b], V.J. Paul[b], G.M. Gremillion[a,c], K. Drnec[a], C. Atwater[b,d], J.R. Estepp[e], J.R. Lukos[f], E.C. Carter[a], and W.D. Nothwang[c]

[a]U.S. ARL– HRED, 459 Mulberry Point Rd., Aberdeen Proving Ground, MD, 21005; [b]U.S. Army TARDEC, 6305 E. 11 Mile Rd., Warren, MI, 48092; [c]U.S. ARL-SEDD, 2800 Powder Mill Rd, Adelphi, MD, 20783; [d]DCS Corporation, 7400 Miller Dr., Warren, MI, 48092; [e]U.S. AFRL, 1864 4th St., Wright-Patterson AFB, OH, 45433; [f]U.S. Navy SPAWAR, 53560 Hull St., San Diego, CA, 92152;

ABSTRACT

All automations must, at some point in their lifecycle, interface with one or more humans. Whether operators, end-users, or bystanders, human responses can determine the perceived utility and acceptance of an automation. It has been long believed that human trust is a primary determinant of human-automation interactions and further presumed that calibrating trust can lead to appropriate choices regarding automation use. However, attempts to improve joint system performance by calibrating trust have not yet provided a generalizable solution. To address this, we identified several factors limiting the direct integration of trust, or metrics thereof, into an active mitigation strategy. The present paper outlines our approach to addressing this important issue, its conceptual underpinnings, and practical challenges encountered in execution. Among the most critical outcomes has been a shift in focus from trust to basic interaction behaviors and their antecedent decisions. This change in focus inspired the development of a testbed and paradigm that was deployed in two experiments of human interactions with driving automation that were executed in an immersive, full-motion simulation environment. Moreover, by integrating a behavior and physiology-based predictor within a novel consequence-based control system, we demonstrated that it is possible to anticipate particular interaction behaviors and influence humans towards more optimal choices about automation use in real time. Importantly, this research provides a fertile foundation for the development and integration of advanced, wearable technologies for sensing and inferring critical state variables for better integration of human elements into otherwise fully autonomous systems.

Keywords: Human-Automation Interaction; Trust in Automation; Decision-Making; Consequence-Based Control; Motion-Based Simulation; EEG; Psychophysiology; Privileged Sensing Framework

1. INTRODUCTION

Automation and autonomous technologies are rapidly becoming a *de facto* part of everyday life for most people in the developed world. Whether as software instantiations (e.g. intelligent search engines, automated sentence completion) or physical embodiments (e.g. home cleaning devices, vehicle safety technology), automations across the spectrum are becoming integrated into technologies that support functional behaviors of people in many contexts, whether personal or professional. A human may be a passenger on an autonomous train, a driver of a car that benefits from automated lane-keeping and collision avoidance, an end-user of a product that was created using a fully automated manufacturing process, or they may simply be an uninvolved bystander in a space where an autonomous agent, such as a home cleaning robot or delivery drone, is operating [1]. Even earlier in the product lifecycle, such as during initial system development or later product testing and evaluation, any automated or autonomous technologies must interact with humans. This recognition that technologies are integrated parts of larger sociotechnical systems has a significant history in the realm of human factors engineering [2,3] and is increasingly relevant to consider more deeply as automation and autonomy become integral to the function and experience of life in human society.

*jason.scott.metcalfe@gmail.com, jason.s.metcalfe2.civ@mail.mil; phone 1 410 278-9229; arl.army.mil

Micro- and Nanotechnology Sensors, Systems, and Applications IX, edited by Thomas George, Achyut K. Dutta, M. Saif Islam, Proc. of SPIE Vol. 10194, 101941U · © 2017 SPIE
CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2264245

An upshot of this discussion is that human responses to technology have always been and will continue to be an important determinant of its perceived utility and safety and thus can both facilitate and undermine how well a given technology is accepted into common use. The degree to which humans accept and appropriately use automated technologies is believed to be heavily influenced by issues of trust and confidence. As the narrative goes, while people gain confidence in the reliability, robustness, and safety of automated technologies, they also develop sufficient trust to willingly share important decision and/or control authority with such systems. The fact that human trust in automation (TiA) is considered to be a major determinant of acceptance and use of these advanced technologies has been well-documented [4-9]. More important than achieving a certain level of trust, however, is understanding how to appropriately match the individual human operator's trust with the relative capabilities of the technology [10-12]. Therefore, an important goal for systems designers has been to identify mechanisms and principles that enable calibrating human TiA to elicit appropriate decisions and behaviors with respect to interacting with automation in the context of an ongoing, dynamic, task context [4,7].

The present discussion describes a collaborative multi-year effort wherein we created a novel research framework for studying real-time changes in operator TiA. The ultimate goal of this work was to develop a tractable, near-term method to enable real-time management of human-automation interaction (HAI) through TiA. The research was part of a three-year program known as the Autonomy Research Pilot Initiative (ARPI) that was funded by the U.S. Office of the Secretary of Defense (OSD). The overall objective of the ARPI was to promote innovative science and technology that would enable autonomous systems to meet future U.S. Department of Defense mission requirements. The TiA research described in this paper was conducted as a specific technical area within a larger ARPI project that aimed to develop a novel conceptual approach for integrating humans and autonomous systems [13]. The following sections outline our approach to building a research framework to address the issue of miscalibrated TiA, the essential concepts and assumptions at its foundation, and practical challenges encountered in execution of the research. Among the most critical outcomes of the new research framework was a shift in focus from attempting to calibrate the individual's psychological trust state towards management of their downstream automation interaction behaviors by influencing their antecedent decisions [4]. This change led to the development of a testbed and paradigm [14] that was deployed in two experiments of human interactions with driving automation, both of which were executed in an immersive, full-motion simulation environment [15,16]. Broader implications of this work touch on technology integration into typical human life, wherein portable computing and wearable sensing are increasingly capable of providing similar kinds of information to facilitate better real-time inference regarding the wearer's state changes throughout the day.

2. BUILDING A NEW RESEARCH FRAMEWORK FOR TIA

The ARPI was conceived by OSD to promote high-risk research that would yield rapid innovations in areas that have been critically challenging to the application of autonomy in operational environments. HAI was identified among these critical challenge areas and miscalibrated TiA, in particular, was expressly called out as a fundamental problem. In response to the broader program announcement, a team from the Army Research Laboratory (ARL), the Air Force Research Laboratory (AFRL), the Army Tank and Automotive Research, Development, and Engineering Center (TARDEC), and the Space and Naval Warfare Systems Center (SPAWAR) adopted an ambitious and high-risk technical approach focused on developing a generalizable set of principles that are broadly applicable to a variety of problems in the domain of HAI and, in this, offered a novel pathway to real-time mitigation of problems typically addressed through TiA calibration.

The entry point for developing a method to calibrate TiA must logically involve access to reliable and objective metrics of how and when it changes. Thus, the initial thrust of the ARPI TiA research involved surveying the state of the art in TiA measurement and then assessing the gaps to be addressed in developing metrics to be incorporated into a real-time system for TiA calibration. As expected, the assessment revealed that subjective self-reports, behavioral inference (e.g. automation use patterns known as reliance or compliance), or a combination of both have been the dominant methods for inferring changes in trust [c.f. 5,17,18]. While providing important insights into factors that influence TiA, these types of metrics were also immediately seen as limited in their ability to facilitate its real-time calibration. Obtaining each self-report measurement would require the operator to shift attention away from their primary task, which would likely be to the detriment of overall performance. Likewise, automation reliance or compliance patterns typically cannot detect changes in TiA until it is too late to use this information in an active calibration strategy (i.e. diminished trust resulting in under-reliance may only be inferred after the inappropriate or detrimental behavior has already happened).

After confirming our suspicion that the state of the art in TiA measurement has been limited to behavioral inference and individual self-reports, we proceeded with looking for real-time methods available elsewhere. A set of methods with which our team has significant experience was drawn from domain of cognitive neuroscience, which offers a host of techniques and metrics that allow real-time, quantitative insight into physiological processes that underlie fluctuating psychological states such as trust. Notionally, we considered the potential of such techniques and metrics because TiA is a subjective state that is very likely to be associated with consistent physiological responses. Ideally, in combination with observable behaviors, physiological metrics would index how each individual was responding to the task context and would provide *advance* indicators of when that state was about to change [4]. If such indicators were to prove useful, then they would also need to be robust within and between individuals as well as across task, time, and technology. Here, we asked whether, through monitoring an operator's physiological state and behavior, objective metrics could be developed to enable effective inference of his or her changing levels of TiA.

Indeed, considerable research on interpersonal trust has revealed measurable patterns of physiological change that correlate significantly with changing levels of subjective trust and trust-based decision making [19]. Encouragingly, methods and measures from psychophysiology have been used to gain insights into other aspects of HAI that are likely to be related to variations in TiA. For instance, despite the reality that few studies have used eye-tracking expressly for trust inference, it has been applied to gain insight into constructs such as visual attention, mental workload, and automation complacency [20,21], each of which have logical and/or empirical connections with TiA. Similarly, with a few relatively recent exceptions [22,23], electrophysiological measurements such as electroencephalogram (EEG), electrocardiogram (ECG), or electrodermal activity (EDA), have not frequently been applied directly to infer trust in automation as a cognitive state. However, as with eye tracking, psycho- and electro-physiological methods have been applied to examine constructs like task engagement and technology acceptance that are believed to have associations with TiA [24]. Moreover, the success of research such as that of Krueger et al. [25] has demonstrated a neural basis for interpersonal trust, which we hypothesized would extend to and be measurable in the context of TiA.

In addition to robust and reliable inference of changing TiA, the notion of its real-time calibration pointed to a need for prediction. In other words, fundamental to the concept of TiA calibration has been a notion that undesirable trust states may be identified while (or even before) they emerge and then be influenced toward a better alignment with the current needs of the task. Indeed, some have already attempted this by predicting aspects of future *system* performance and *environmental* states (e.g. the increasing likelihood of a vehicle collision) and then issuing feedback (e.g. warnings) to inform and enhance operator trust in the system as well as affect appropriate reliance or compliance behavior [12,26]. While somewhat effective when used appropriately, this type of method may also be too indirect and too reliant on assumptions that the operator would consistently use the information in predictable ways and then make appropriate decisions regarding how to interact with the automation. Fundamentally, building calibration strategies on such notions would mean defining feedback functions based on average performance across many individual operators observed under a limited set of commonly-experienced conditions rather than tailoring to a specific human operator across a broad variety of conditions.

Given known inter- and intra-individual variability in how psychological states are manifest in behavior as well as which specific conditions provoke those manifestations, we considered the above notions of trust-calibration as risking being only marginally effective, in the best case, and counter-productive at worst. That is, for those who behave in accordance with the design assumptions under anticipated conditions, the system would likely be judged as trustworthy and be used appropriately. However, for those who vary from assumed behaviors or who experience circumstances for which the automation was not explicitly tuned, the system would likely fail and end up being judged as unreliable and untrustworthy. We thus pursued an alternative that was based on direct, real-time measurement of the human, the automation, and the environment, and then sought to develop an *a priori* understanding of individualized predictors of likely trust-related behaviors such as over- and under-reliance on the automation. By making a shift away from using trust as a controlled variable and, instead, towards predicting and influencing resultant behaviors, we also viewed this as a move away from nominal concepts of trust calibration and more firmly towards management of trust-based behaviors. The section that follows provides a summary of concepts that were critical enablers of this approach.

3. CRITICAL CONCEPTS

Figure 1 provides a high-level diagram of the initial concept behind the ARPI TiA research project. Shown here is a hypothetical circumstance where an operator remotely supervises and manages the behavior of a semi-autonomous asset. Using online physiological and performance monitoring in combination with context information and weighting functions that would be determined in *a priori* fashion, a decision would be made to influence the feedback to the operator, the behavior of the agent, or both. The control feedback in this case would be determined based on the principles outlined in the privileged sensing framework [13], which include (a) confidence metrics to account for the variance in quality of estimates from each information stream about each agent, (b) estimated reward relative to risk of providing feedback or assigning control to each agent, and (c) integrated information from data streams operating on different temporal, spatial, and decision or estimation scales.

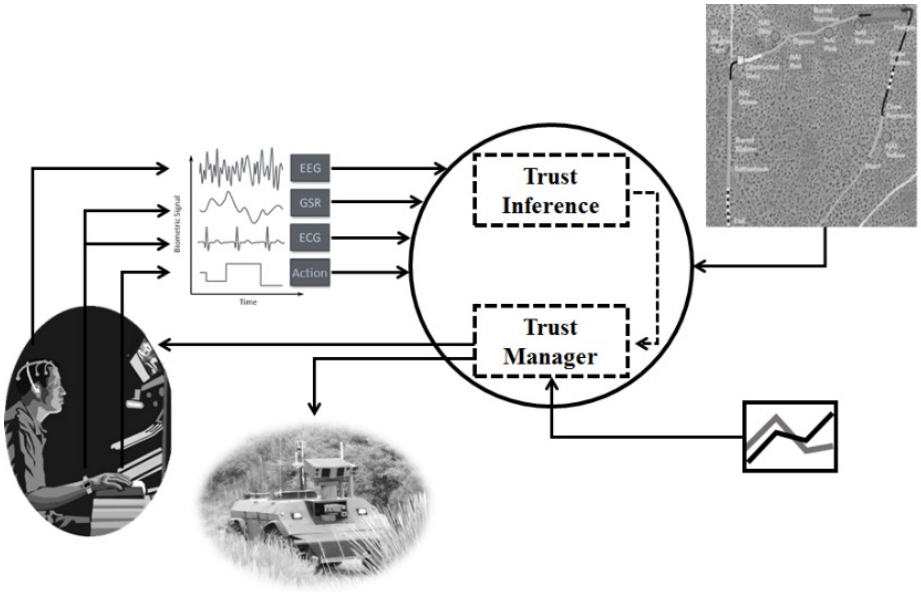

Figure 1. Initial concept for a real-time TiA management system. Starting at the lower left is a remote operator of an autonomous asset, who would be outfitted with various lightweight physiological sensors. Physiological and behavioral data would be processed within a Privileged Sensing Framework based controller (see text and [13]) to infer the current operator "trust state". Taken in combination with task inputs (e.g. map data, top right) and consequence-based privilege functions (e.g. inset curves, bottom right) that would be determined based on *a priori* data, a trust manager would decide whether to provide feedback to the operator or execute a corrective action in the asset.

As applied to the TiA project, the first objective was to develop the inner workings of the state estimator (Figure 1, "trust inference") and then, later, determine a control method to enact mitigations for miscalibrated trust (Figure 1, "trust manager"). Given that state of the art methods in TiA inference have mainly been based on *a posteriori* behavior and subjective self-reports, we considered it necessary to shift the focus of our effort in a more practical direction that could address the same problem space while enabling a near-term solution for real-time management of the ongoing HAI. Instead of attempting to directly measure and manage TiA, the ARPI research focused on mitigation strategies at the level of operator decisions about behaviors. This was necessary because the relationship between TiA and interaction decisions is complex and, more importantly, the extant literature has yet to agree on any specific predictive relationship between TiA and behavior that might reasonably be implemented in real-time control for management of HAI in the near future [4].

Owing to the decision to shift focus towards direct behavioral mitigation at the level of decision-making, the second objective of the research was to establish a method for incorporating predictions about trust-based decision-making into an active decision management scheme. In particular, we operationalized "trust inference" using an algorithm to predict impending operator decisions and subsequently, "trust management" was operationalized through control logic that dictated instantaneous feedback that was given to the operator regarding which decision might optimize the likelihood of error-free performance. Focusing on the operator decisions about transitions in control authority (i.e. reliance) provided

an objective behavior that could be leveraged for the management component, even if not solely or directly based on TiA.

4. IMPLEMENTATION AND SOME LESSONS LEARNED

This section describes, in more specific terms, how the concepts described above were translated into a testbed paradigm and set of experiments. The discussion provides a high level summary of the results of the research and discusses a few of the more important lessons that were learned along the way to understanding how to better leverage human variability to facilitate enhanced operator TiA. As elements of this research have been published elsewhere, only a summary of the relevant aspects of the paradigm and data are provided here. For more complete details on the research, please refer to the previous publications [14-16] as well as successive works from the same authors (in preparation for near future publication).

4.1 From concepts to experimental methods and observations

Figure 2 provides a visual summary of the apparatus and paradigm that constituted the ARPI-developed testbed for research in active management of HAI through understanding TiA and its dynamics. The research leveraged a 6 degree-of-freedom, full-motion driving simulator in the Ground Vehicle Simulation Laboratory at the U.S. Army TARDEC facilities in Warren, MI. Driving was chosen because it is a common task with which most adults in the U.S. are familiar and thus required minimal advance participant training. Moreover, civilian vehicles are increasingly becoming targets for implementation of various types of automation and autonomous technologies and therefore, understanding how drivers develop trust in a vehicle-based automation is a timely and relevant topic of research.

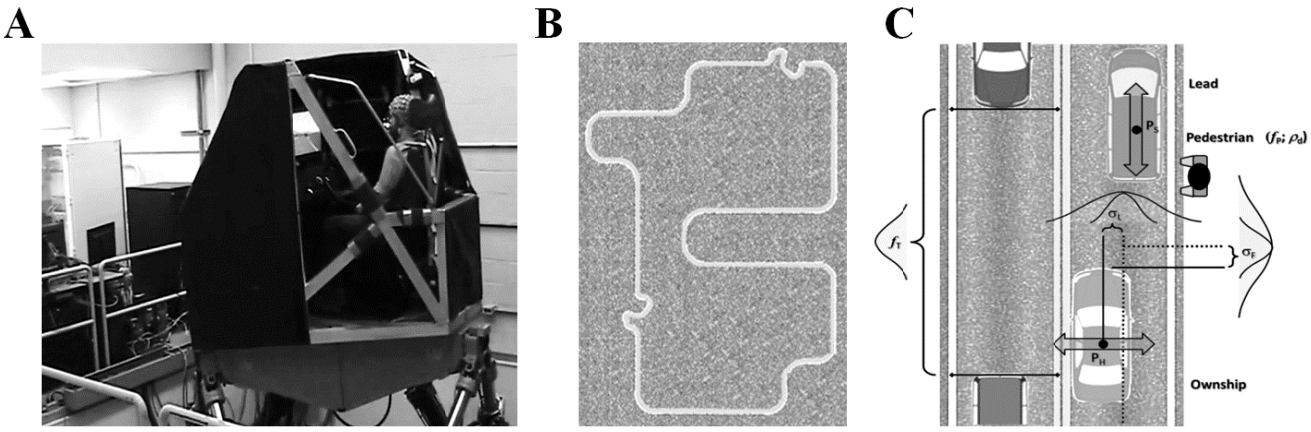

Figure 2. Depictions of the apparatus and paradigm for the ARPI TiA research. (A) Full-motion driving simulator used for the ARPI TiA research at TARDEC facilities in Warren, MI. (B) Map of the course on which the simulated drives occurred. (C) Summary depiction of the leader-follower paradigm that was developed to study human interactions with driving automation in conditions that were intended to influence TiA. The car labeled "**ownship**" represents the vehicle controlled by the participants and the car labeled "**lead**" represents the target vehicle to be followed. $\mathbf{P_H}$ refers to the heading perturbations (wind gusts) and $\mathbf{P_S}$ refers to speed perturbations. σ_L and σ_F indicate the variance in automation performance at lane keeping and speed following, with wider distributions indicating lower quality automation performance. f_T indicates a variable frequency of oncoming traffic. f_P specifies the overall appearance frequency of pedestrians and ρ_d indicates the proportion of those pedestrians that were dynamic, meaning they would walk in the path of the vehicle.

The ARPI testbed paradigm that was used for all of the TiA research was based on a leader-follower driving task. Participants were asked to drive around a simulated roadway while maintaining a distance within a set range of a lead vehicle and staying in their lane without colliding with other vehicles or pedestrians. Throughout each session, driving conditions were challenged by pseudo-randomized perturbations where either the lead vehicle would suddenly change speed (up or down) or a wind gust would push the driver's vehicle towards (or even out of) one of the lane boundaries. It was never the case that both types of perturbation would co-occur simultaneously. In addition to driving, participants were instructed to use a game-controller and identify pedestrians that appeared in front of the vehicle, randomly to either

the left or right side of the road. A randomly-chosen 15% of the pedestrians would begin walking into the path of the vehicle after appearing. Participants were instructed to use the game controller to "clear" the pedestrian from their path by pressing an appropriate button (one button for standing still and a different button for walking pedestrians); incorrect button presses would result in a failure to clear the pedestrian out of the way, though a fast enough secondary reaction could correct the error. For both experiments, the driving task was gamified by using a point scheme that set a hierarchical risk structure. Pedestrian collisions cost the most (-100 pts), then other vehicles (-50 pts), followed by minor penalties for deviating outside of the task parameters (i.e. stay in lane and following range, -2 pts per 3 s in error, and pressing incorrect buttons for pedestrians, -5 pts). The final points later translated into a bonus payment received at the end of each experiment and participants knew that each error was effectively going to cost money out of their bonus. This definite consequence structure later factored into the development of the "trust inference" and "trust management" algorithms (Figure 1).

Across two different studies using this testbed, participants were provided access to several different driving automations, putatively to assist their task performance as they saw fit. The automations chosen for these experiments included lane-keeping and adaptive cruise control, but no collision avoidance or advance braking. All participants in each experiment provided written informed consent in accordance with procedures approved by the Institutional Review Board of the US Army Research Laboratory and all testing conformed to the guidelines set forth by the1964 Declaration of Helsinki.

In the first experiment, the participants performed five conditions that included a manual driving baseline and then two using full automation (lane-keeping with speed control) and two using speed control only; in each of the automation type subsets (full, speed-only), there was a high-quality and low-quality variant (see distributions in Figure 1). Therefore, the four conditions in which a driving automation was available included full-high (FH), full-low (FL), speed-high (SH), and speed-low (SL). The intention of this 2 x 2 arrangement (type x quality) was to differentially influence driver trust while controlling for potentially confounding effects of task load (aka workload). Per design, a general characterization of the task load, from lowest to highest, was $FH \leq SH \simeq SL < FL$. This task load differential was controlled through increasing the rate and strength of wind-gust perturbations in SH in order to equalize the overall effort the participants had to expend to maintain driving performance in each condition. Nothing was done to expressly offset the task load in FH versus FL conditions and this allowed potential TiA comparisons under changing task load as well as when it was approximately equal (in SL/SH). The experimental manipulation of task load, and its impact on subjective workload, was later validated by participant responses on a standard subjective assessment [14] that showed a significant statistical interaction where the FL condition was associated with the highest workload and all other automation conditions were approximately equivalent (the manual control baseline condition was associated with the highest subjective workload of all).

The first experiment established ground truth regarding the conditions most likely to provoke control authority transitions and agent-specific (human vs auto) error rates. By examining driving performance as a function of task conditions (e.g. perturbations) as well as course feature (e.g. straight versus curved segments), it was established that humans had certain competencies that the driving automation did not and likewise, there were conditions under which the automation was more competent. For instance, the driving automation had a tendency to cut curves short, likely because of how the auto navigation algorithm was designed as a variant of waypoint-like path specification. This was observed in the first experiment as a strong tendency of the automated driver in the FH and FL conditions to have greater rates of lane violation errors than humans in segments with increasing curvature as compared with straight segments. Human drivers, on the other hand, tended to perform more poorly than the driving automation under longitudinal perturbation conditions (i.e. speed changes). In part, this was likely a consequence of humans tending to more readily detect differential self-versus-other velocities in their peripheral vision (such as a passing vehicle) than in their central visual field (such as speed changes in a vehicle driving directly in front of them). In general, throughout the experiment, the participants patterns of reliance tended to reflect these relative competencies as well as the general quality of the automation. That is, they tended to be more willing to use the speed-control automation when it was highly reliable (SH), but were much more variable in their willingness to rely on the same automation when its performance quality was reduced (SL). Likewise, the human participants were almost always *unwilling* to rely on the lane-keeping automation during the tight "s-curve" segments regardless of automation quality, but showed some increased tendency to allow the high quality lane-keeping automation (FH) to navigate more gradual curves as compared with the FL condition, potentially indicating greater TiA [see 15 for more detailed analyses].

The second experiment focused on developing a better understanding of how human drivers would respond to and use information that was provided to encourage selection of optimal reliance behaviors. The second study leveraged the ground truth that was described above into two different control architectures that were designed to influence reliance behaviors in real time. Unlike the previous experiment, there was only one automation available to the participants. The performance of the available automation was comparable to the FH condition from the first study; the automation provided lane-keeping and adaptive speed control with performance characteristics that were similar to, but not as good as, the FH automation from the first experiment. The reason for selecting this automation was to create specific opportunities in which the human driver was superior to the automation and vice-versa, such that the preferred mode could be experimentally controlled.

The overall driving task remained the same for this second experiment, including the leader-follower structure, driving perturbations, and pedestrians. As with the first experiment, all participants performed a manual baseline drive prior to driving with the automations. The data recorded in this baseline provided an individualized set of information upon which relative risk between the driver and automation could be calculated; secondarily, but nonetheless important, this also provided a means of accounting for inter-individual variability. Then in three successive conditions, participants completed drives with the automation present. On one of these experimental drives, participants were not provided recommendations as to which control mode to use at any particular time. This served as a replication condition for comparison across the two studies and provided data sufficient to characterize each participants' individual preferences in reliance behavior. Then, in two other conditions, the participants completed experimental drives in which a feedback control model calculated real-time predictions of optimal reliance behavior and then provided on-screen feedback that gave a continuously graded suggestion of control mode (manual vs. auto) such that the strength of the preference was indicated as well as the direction. The two feedback conditions differed in a detail that enabled testing of a principle of the privileged sensing framework, but this detail does not merit discussion here. See [16] for a more complete report of the conditions and how they influenced the control architecture.

Overall performance on the driving task remained consistent with the ground truth data from the first experiment. Moreover, use of the automation in any condition generally improved performance over the manual driving baseline, which was also true in the first experiment. Initial analyses have suggested that use of the control feedback led to a general improvement in performance as revealed in a lower likelihood of driving errors compared with error rates when the participants chose to operate in a mode other than suggested by the on screen feedback. As compared with the automation condition for which no control suggestions were provided, there was mild evidence for a change in the participants' decisions regarding when to rely on the available automation. While preliminary, the observed pattern of change suggests that the participants' decisions regarding when and how to rely on the automation were closer to optimal than when no such feedback was provided.

Taken together, the initial observations from the first two experiments provided encouraging evidence that the conceptual approach described in section 3 may have provided an important, practical, and near-term implementable strategy for mitigating HAI problems in the context of dynamically-varying TiA. Specifically, by developing an understanding of the risk structure of the task given probable performance outcomes of both agents, it was possible to develop an objective function upon which a real-time control strategy was based. Moreover, using that information and finding methods to adequately facilitate operator adoption of recommendations may have led to improved overall joint human-automation performance. Finally, thus far the effects of the mitigation strategy have been moderate and this suggests that the simple provision of visual information is not likely to be sufficiently influential, but that there may be circumstances under which directly enforcing a particular change in control authority would be warranted. How to take such a step so as to positively influence TiA without undermining it, of course, remains an open challenge.

4.2 Challenges and lessons

Beyond the initial results, there is value in reporting on some of the most significant challenges that were encountered in building and implementing this research framework. As has been emphasized thus far, our framework considers risk and consequence as fundamentally important to the study of HAI in general as well as TiA in particular. Therefore, an important early component of the research framework necessarily required establishing a clear risk structure and, consequently, defining circumstances under which interaction behaviors (specifically, authority change decisions) would vary in systematic ways. Without achieving this, the research described above would not have allowed for inference of causal inferences about how humans make choices about reliance on an available automation. In the efforts to account

for this concern, we learned an important lesson involving the eventual recognition of how perceived risk may actually factor into automation reliance decisions. That is, through building the research framework, we developed a new understanding of how human operator decisions are affected by perceived relative capability and, further, how individual biases may play a fundamental role in the formation of that perception.

On first consideration, the statement above may seem relatively intuitive, especially to those who understand the literature in this domain. There have been a variety of experiments establishing that people tend to trust automations that will outperform them at a given task. Conversely, people tend to lack trust in an automation that makes performance errors that they judge they would not make [27]. Neither of these statements is particularly controversial. A self-interested agent will tend to use the resource most likely to help them achieve their goal. Considering this conventional wisdom, our initial settings for automated driving performance in the low and high performance conditions seemed relatively straightforward; the "trusted" (high) conditions involved making the automation clearly better at driving than humans and the "untrusted" conditions involved making the automation clearly worse than humans. However, it was quickly observed in pilot testing that the automation that was clearly worse than the participants was completely distrusted and unlikely to be considered useful at all, resulting in what Parasuraman and Riley have referred to as disuse [9]. Yet, disuse was not the intended object of study. Instead, the research framework was designed to learn how to more precisely correct miscalibrated TiA by gradually lowering it when over-trust was observed and, reciprocally, gradually increasing TiA in the face of under-trust.

Therefore, it was important to create lower quality automations that were *less* trustworthy but still considered occasionally useful by the participants. This was eventually achieved by setting the performance of the low quality automations to statistically replicate another human. Specifically, the low-performing automation actually shared a number of basic driving characteristics (such as lane and speed variance) with the average driving performance observed in a small sample of pilot participants who completed the task manually. Though not a formal experimental result, we took this as a clear lesson. An average human may be more willing to trust an entity that behaves like another average human, but because of individual biases due to awareness that the other agent is automated rather than human, a small reduction of trust may be induced without incurring complete distrust. Our initial results suggested that when we used a design for the lower quality automation that leveraged this knowledge, we observed no participants that fell into a pattern of disuse due to a complete loss of TiA. This provided an important insight into how to gradually correct over-trust conditions without causing a complete loss in trust. Of course, further research will be needed to fully understand this anecdotal conclusion, but such an insight may be an important key to eventual real-time calibration or trust-based decisions regarding automation reliance.

A second major challenge in building the TiA research framework was faced when considering methods to influence individual decisions in real-time without directly enforcing a specific decision by forcibly removing authority from or imposing authority on the driver. That is, because we have observed people distrust agents when they believed that they did not have sufficient control [28], it was considered important to enact control optimization by encouraging the participants to make any changes themselves at any time that they preferred. Yet, even with adequate prediction of the preferred control mode (i.e. the one most likely to lead to better performance), a problem remained as to how to reliably elicit the selection of that mode by the human. This was especially challenging given the known problem of individual biases. Though it was suspected to be a weak influence, a visual indicator was ultimately chosen as the mechanism of influence over the participants. The main concept was to provide transparency to the participants such that they would be able to apprehend when recommendations were particularly strong as well as when the preference represented only a weak advantage for one mode over the other. Ideally, stronger preferences would have been more likely to encourage the participants to comply with the recommendation and the weaker preferences would have allowed for more individual choice, thus being useful and at the same time enabling a sense of agency to accommodate individual preferences. Notionally, allowing the participants to remain an active and essential element in the control loop was intended to also improve their TiA while facilitating optimal performance.

As described in the previous section, the data that have been examined thus far provide a small degree of support for the potential effectiveness of the approach, but it remains clear that significant room for improvement continues to exist. From the efforts expended thus far, the biggest challenge appears to reside in countering the individual biases. Participants were selective about when they chose to follow the control recommendations and, as indicated above, this limited the effectiveness of the mitigation scheme. It is likely that a secondary method for influencing decisions will be necessary.

4.3 Opportunities

In the second experiment, we attempted to enact real-time mitigation solely with predictive individualized feedback to the operator. However, as shown in Figure 1, another option may be to change some aspect of the performance of the automation instead. This may be especially attractive in cases where it seems that the operator ignores the feedback given to them. Indeed, cognitive biases leading to neglect of relevant information have long been understood to be problematic in HAI [9]. For example, when using increasingly capable automations, some operators experience phenomena such as "out of the loop syndrome" and "automation-induced complacency" in which they become passive with respect to engaging the automation and task [29,30]. Likewise, in tasks that involve a relatively stable visual flow, people can fall prey to a phenomenon known as "attentional tunneling", which also impacts the degree of their active engagement in the task [31]. It is possible to detect when an operator is in such states by implementing monitoring methods that are capable of indexing changes in associated physiological variables [4,32]. Reliable detection of such states then can create an opportunity to define mitigations and this is where we assert an opportunity exists to enact a change by affecting the behavior of the automation rather than acting through simple visual feedback to the operator.

In both of the experiments described in this paper, participants were outfitted with an array of physiological monitors that provided insight into brain and bodily indicators of cognitive state. In recent analyses of these data, we have begun to observe relationships between the physiological variables and operator decisions to rely or not on the automation. Our early results have been particularly encouraging with respect to identifying of instances of impending driver take-over from the automation by using the physiological monitoring. Should these results hold, they may enable addressing onset of automation complacency, or attentional tunneling, during instances in which indicators of impending take-over decisions are absent. In such a case, for instance, a command could be sent to the automation to reduce throttle by a small but safe amount and this may be enough bring the driver's attention back into the loop. Indeed, during the first experiment we observed that lead vehicle speed changes were reliable triggers of a driver take-over, especially in the SL condition.

5. CONCLUSION

This paper describes a successful research program that led to concrete methods to address the key problems identified in the space of miscalibrated TiA. Observation of performance of the human and automated agents under different task conditions enabled a means of understanding the probable risk of operating in different control regimes (human vs. auto) given ambient task demands. This method also enabled integrating confidence-based decision making as a function of the relative risk and reward predicted from *a priori* observations of each agent's performance under similar task constraints. Direct behavioral and wearable physiological monitoring provides a continuing opportunity to account for individual variability in the relationship between the specific human operator and the driving automation. Leveraging these sources of information within a unified research framework has enabled a practical approach to real-time management of HAI in the context of dynamically varying TiA and we believe that this provides an example and creates an opportunity to more fully understand how to integrate humans with everyday automation and other advanced technologies.

The overall purpose of this paper was to describe our recent efforts to develop and execute a novel scientific approach to address long-standing challenges in the domain of HAI. Here, the focus was on the development of a research framework, a novel paradigm, and a system design approach to enable near-term solutions for the problems that are commonly associated with miscalibrated TiA. Over the three-year program funded by OSD, our deep analysis of the HAI problem space through extensive literature review combined with the design and execution of novel experiments intended to identify essential behavioral and physiological indicators of operator performance decisions, such as the studies described herein, have provided important evidence for consideration of alternative ways to incorporate humans with automated systems that leverage modern wearable sensing technology. This paper has highlighted the significant efforts that were required to address this problem space and some of the early successes and opportunities for future progress that have resulted from those efforts. We believe that similar efforts will needed to more fully address the broad spectrum of challenges in HAI as automation and advanced technologies increasingly pervade daily life.

ACKNOWLEDGEMENTS

The research described in this paper was supported by the Office of the Secretary of Defense through the Autonomy Research Pilot Initiative under MIPR DWAM31168. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] Scholtz, J., "Theory and evaluation of human robot interactions," Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03), (2003).
[2] Carayon, P., "Human factors of complex sociotechnical systems," Applied Ergonomics 37, 525-535 (2006).
[3] Walker, G.H., Stanton, N.A., Salmon, P.M., and Jenkins, D.P., "A review of sociotechnical systems theory: A classic concept for new command and control paradigms," Theoretical Issues in Ergonomics Science 9(6), 479-499 (2008).
[4] Drnec, K., Marathe, A.R., Lukos, J.R., and Metcalfe, J.S., "From Trust in Automation to Decision Neuroscience: Applying Cognitive Neuroscience Methods to Understand and Improve Interaction Decisions Involved in Human Automation Interaction," Frontiers in Human Neuroscience 10, 290 (2016).
[5] Dzindolet, M., Peterson, S., Pomranky, T., and Beck, P., "The role of trust in automation reliance," International Journal of Human-Computer Studies 58(6), 21 (2003).
[6] Lee, J. and Moray, N., "Trust, control strategies and allocation of function in human-machine systems," Ergonomics 35(10), 1243-1270 (1992).
[7] Lee, J.D. & See, K.A., "Trust in automation: Designing for appropriate reliance," Human Factors 46(1), 50-80 (2004).
[8] Muir, B. M., "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," Ergonomics 37(11), 1905-1922 (1994).
[9] Parasuraman, R. and Riley, V., "Humans and automation; use misuse, disuse, abuse," Human Factors, 39(2), 23 (1997).
[10] McGuirl, J. M. and Sarter, N. B., "Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information," Human Factors 48(4), 656-665 (2006).
[11] Merritt, S. M., Huber, K., LaChapell-Unnerstall, J., and Lee, D., "Continuous Calibration of Trust in Automated Systems", Air Force Research Laboratory Technical Report AFRL-RH-WP-TR-2014-0026, (2014).
[12] Cai, H. and Lin, Y., "Tuning trust using cognitive cues for better human-machine collaboration," Proceedings of the Human Factors and Ergonomics Society Annual Meeting 54, 2437-2441 (2010).
[13] Marathe, A.R., Metcalfe, J.S., Lance, B.J., Lukos, J., Jangraw, D., Lai, K.T., Touryan, J., Stump, E., Sadler, B.M., Nothwang, W., and McDowell, K., "The privileged sensing framework: A principled approach to improve human-autonomy integration," Theoretical Issues in Ergonomics Science, (in press).
[14] Drnec, K. and Metcalfe, J.S., "Paradigm development for identifying and validating indicators of trust in automation in the operational environment of human automation integration," Lecture Notes in Computer Science (Foundations of Augmented Cognition) 9744 (2016).
[15] Gremillion, G.M., Metcalfe, J.S., Marathe, A.R., Paul, V.J., Christensen, J., Drnec, K., Haynes, B., and Atwater, C., "Analysis of trust in automation for convoy operations," Proc. SPIE 9836 (2016).
[16] Nothwang, W., Gremillion, G.M., Donavanik, D., Haynes, B.A., Atwater, C.S., Canady, J.D., Metcalfe, J., and Marathe, A., "Multi-sensor fusion architecture for human-autonomy teaming," Resilience Week, (2016).
[17] Lyons, J.B., Stokes, C.K., Eschlemen, K.J., Alarcon, G.M., and Barelka, A.J., "Trustworthiness and IT suspicion: An evaluation of the nomological network," Human Factors 53(3), 219-229 (2011).
[18] Moray, N., Inagaki, T., and Itoh, M., "Adaptive automation, trust, and self-confidence in fault management of time-critical tasks," Journal of Experimental Psychology: Applied 6(1), 44-58 (2000).
[19] Borum, R., [The science of interpersonal trust], The MITRE Corporation , McLean, VA (2010).
[20] Metzger, U., Duley, J.A., Abbas, R., and Parasuraman, R., "Effects of variable-priority training on automation-related complacency: Performance and eye movements," Proceedings of the IEA/HFES Congress, 349-349 (2000).

[21] Rovira, E. and Parasuraman, R., "Transitioning to future air traffic management: Effects of imperfect automation on controller attention and performance," Human Factors 52(3), 411-425 (2010).
[22] Bethel, C.L., Salomon, K., Murphy, R.R., and Burke, J.L., "Survey of psychophysiology measurements applied to human-robot interaction," The 16th IEEE International Symposium on Robot and Human Interactive Communication, (2007).
[23] Montague, E., Xu, J., and Chiou, E., "Shared experiences of technology and trust: An experimental study of physiological compliance between active and passive users in technology-mediated collaborative encounters," IEEE Transactions on Human-Machine Systems 44(5), 614-624 (2014).
[24] Fairclough, S.H. and Venables, L., "Prediction of subjective states from psychophysiology: A multivariate approach," Biological Psychology 71, 100-110 (2006).
[25] Krueger, F., McCabe, K., Moll, J., Kriegskorte, N., Zahn, R., Strenziok, M., Heinecke, A., and Grafman, J., "Neural correlates of trust," Proceedings of the National Academy of Sciences USA 104, 20085-20089 (2007).
[26] Gu, W., Mittu, R., Marble, J., Taylor, G., Sibley, C., Coyne, J, and Lawless, W.F., "Towards modeling the behavior of autonomous systems and humans for trusted operations," 2014 AAAI Spring Symposium Series, (2014).
[27] Madhavan, P., Weigmann, D.A., and Lacson, F.C., "Automation failures on tasks easily performed by operators undermine trust in automated aids," Human Factors 48(2), 241-256 (2006).
[28] Metcalfe, J.S., Alban, J., Cosenzo, K., Johnson, T., and Capstick, E, "Field testing of teleoperation versus shared and traded control for military assets: An evaluation involving real-time embedded simulation and Soldier assessment," Proc SPIE 7692, (2010).
[29] Endsley, M.R., Bolté, B., and Jones, D.G. "SA demons: The enemies of situation awareness," [Designing for Situation Awareness: An Approach to User-Centered Design], Taylor & Francis, London, 31-42 (2003).
[30] Bailey, N.R., and Scerbo, M.W., "Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust," Theoretical Issues in Ergonomics Science 8(4), 321-348 (2007).
[31] Wickens, C.D. and Alexander, A.L., "Attentional tunneling and task management in synthetic vision displays," The International Journal of Aviation Psychology 19(2), 182-199 (2009).
[32] Régis, N., Dehais, F, Rachelson, E., Thooris, C., Pizziol, S., Causse, M., and Tessier, C., "Formal detection of attentional tunneling in human operator-automation interactions", IEEE Transactions on Human-Machine Systems 44(3), 326-336 (2014).