Hacking the Human: The Prevalence Paradox in Cybersecurity

Article *in* Human Factors The Journal of the Human Factors and Ergonomics Society · August 2018

DOI: 10.1177/0018720818780472

2 authors:

Ben D Sawyer
Massachusetts Institute of Technology
33 PUBLICATIONS **186** CITATIONS

SEE PROFILE

Peter A Hancock
University of Central Florida
472 PUBLICATIONS **8,398** CITATIONS

SEE PROFILE

SPECIAL SECTION: 2017 **Human Factors** *Prize Winner*

Hacking the Human: The Prevalence Paradox in Cybersecurity

Ben D. Sawyer, Massachusetts Institute of Technology, Cambridge, and **Peter A. Hancock**, University of Central Florida, Orlando

Objective: This work assesses the efficacy of the "prevalence effect" as a form of cyberattack in human-automation teaming, using an email task.

Background: Under the prevalence effect, rare signals are more difficult to detect, even when taking into account their proportionally low occurrence. This decline represents diminished human capability to both detect and respond. As signal probability (SP) approaches zero, accuracy exhibits logarithmic decay. Cybersecurity, a context in which the environment is entirely artificial, provides an opportunity to manufacture conditions enhancing or degrading human performance, such as prevalence effects. Email cybersecurity prevalence effects have not previously been demonstrated, nor intentionally manipulated.

Method: The Email Testbed (ET) provides a simulation of a clerical email work involving messages containing sensitive personal information. Using the ET, participants were presented with 300 email interactions and received cyberattacks at rates of either 1%, 5%, or 20%.

Results: Results demonstrated the existence and power of prevalence effects in email cybersecurity. Attacks delivered at a rate of 1% were significantly more likely to succeed, and the overall pattern of accuracy across declining SP exhibited logarithmic decay.

Application: These findings suggest a "prevalence paradox" within human-machine teams. As automation reduces attack SP, the human operator becomes increasingly likely to fail in detecting and reporting attacks that remain. In the cyber realm, the potential to artificially inflict this state on adversaries, hacking the human operator rather than algorithmic defense, is considered. Specific and general information security design countermeasures are offered.

Keywords: human-computer interaction, internet, information security, messages, signal detection, vigilance, risk, antivirus, virus, antimalware, malware, design

Address correspondence to Dr. Ben D. Sawyer, Department of Engineering, AgeLab, Massachusetts Institute of Technology, 1 Amherst St. Cambridge, MA 02139, USA; e-mail: bsawyer@mit.edu or sawyer@inhumanfactors.com.

HUMAN FACTORS
Vol. 60, No. 5, August 2018, pp. 597–609
DOI: 10.1177/0018720818780472

Cyberattack represents one of the most destabilizing global effects to the Anthropocene technical infrastructure that frames our lives today. Annual global costs are presently a remarkable $500 billion and are projected to quadruple by 2019 (Moar, 2017). Cyberattack and concomitant cyberdefense efforts are heavily, and in many cases even exclusively, vested within algorithmic and software realms. However, cybersecurity is fundamentally a case of human-automation teaming and one in which both the machine and human are potentially vulnerable. Consider the extensive ecosystem of autonomous assistance involved in simply checking email, an act that presently delivers more malware than any other digital vector (Symantec, 2016). Antimalware stops arbitrary code execution, spam detection diverts social engineering attempts, and informational agents choose opportune moments to suggest digital hygiene (as in Sawyer et al., 2015). Users, so well-shielded, may encounter unfiltered attacks only very rarely. Indeed, such successful algorithmic protection may invite "the prevalence effect," in which as signals become less common, they become substantially more difficult for humans to detect. Such diminished acuity underscores a fundamental human inability to detect and respond to the extremely rare signals, however critical they are (Hancock, 2013; Hancock & Warm, 1989; Warm & Jerison, 1984; Wolfe, Horowitz, & Kenner, 2005). When machine successes become the seeds of human failure, what implications to human-machine teaming arise from this "prevalence paradox"? If such conditions can be elicited by an enemy, might this vulnerability of human cognition be weaponized (as suggested in Sawyer et al., 2016)?

Prevalence effects in cybersecurity have previously been experimentally identified only in the context of dedicated teams of military cyberdefenders working with internet traffic "waterfall displays" (Sawyer, Finomore, Funke, & Warm, 2014; Sawyer et al., 2016). Such work

has joined a number of diverse applied visual search scenarios implicating the prevalence effect. From baggage and radiological screening to decisions to employ lethal force (Schultz, Matthews, Warm, & Washburn, 2009; Wolfe et al., 2005), when signals of interest are rare, compared to nonsignal events, human observers are far more likely to simply fail to respond (Hancock, 2013). A key arbiter of the effect is the ratio of signals compared to total events to be evaluated, referred to as "signal probability" (SP). Work plotting hundreds of SPs across millions of trials has revealed that the pattern of decrease of probability to respond exhibits not a linear change but rather logarithmic decay (Mitroff & Biggs, 2014). Practically speaking, there exists a logarithmic "tipping point" after which performance declines very rapidly. This log pattern means that the SP threshold that differentiates between robust detection and escalating failure can be very fine. Interestingly, given the opportunity to correct shortfalls, compromised observers reconsider and respond to the majority of missed signals (Fleck & Mitroff, 2007), inviting the possibility that the prevalence effect cripples not the ability to detect but rather the ability to act. While in tasks such as radiological detection of tumors there exists the opportunity to check upon initial suppositions, perhaps even several times, in many applied contexts failures to act in a timely manner are tragically uncorrectable.

Real-world signals of critical import are often extremely rare, and the rate of their future prevalence is also unknown and uncontrollable. This helplessness on the part of the observer induces the problem characterized as "hours of boredom and moments of terror," or even "months of monotony followed by milliseconds of mayhem," a performance profile shared by many domains (Hancock, 1997). For example, before an improvised explosive device (IED) is encountered, many uneventful deployments in enemy territory may occur (Szalma, Schmidt, Teo, & Hancock, 2014); an unheralded decision about a malicious file may come after hours of email work have passed. Individuals engaged in such tasks report high workload (Finomore, Shaw, Warm, Matthews, & Boles, 2013; Warm, Parasuraman, & Matthews, 2008) generally expressed in the mental demand and frustration subscales of the commonly used NASA TLX (Hart & Staveland, 1988). Victims of SP depression become far more likely to simply fail to respond, allowing malicious information to penetrate their system.

From the perspective of a cyberattacker, the potential to elicit such difficulties in an adversary is indeed alluring. Indeed, artificial SP depression meets the bar for the military concept of "negation," in which active measures are taken to deceive, disrupt, degrade, deny, or destroy opposing capabilities (as in Maybury, 2012; see also Hancock, 2015). From the perspective of the defender, cyberspace has advantages as well. Cyberdefense is a context in which the environment is entirely a human creation and in which any sensory representation of the space may be tailored to human perception by interface designers. While it is difficult to imagine changing the realities surrounding the radiological detection of cancer or the chains of coincidence surrounding friendly fire, in cyberdefense many more degrees of freedom remain available for active intervention. Interestingly, this approach falls short of the definition of social engineering in the context of information security (as in Anderson, 2008) and may constitute a new category of cyberattack. By any name, intentional elicitation of SP depression seems likely to be a versatile, multicontext attack that could be deployed in cyber contexts as diverse as masking backdoors in software, disguising malicious network traffic, or eliciting users to click on email-delivered cyberattacks.

The primacy of email as a digital communication channel renders it a default target for cyberattack. Email is the primary vector for malware propagation, and the frequency of email-delivered cyberattack doubled between 2014 and 2016 alone. Interestingly, in the same period, phishing attacks (email attacks inducing individuals to improperly provide personal information) declined to only 37% of their 2014 levels (Symantec, 2016). The impact of phishing was certainly not in decline. A phishing cyberattack sent to John Podesta, the 2016 Clinton presidential campaign's chairman, resulted in a stolen trove of damaging emails being published online. Similar attacks against other Democratic National Committee (DNC) personnel, as well as against political think tanks, made phishing a

feature of 2016 Presidential election news. More recently, Google users were targeted by widespread and successful email-delivered phishing attacks compromising thousands of accounts (*Washington Post*, 2017). In fact, in 2016, phishing and delivery of malicious code accounted for 40% of total cyberattack costs to companies (Ponemon Institute, 2016). So amidst greater impact, why are the absolute numbers of phishing attacks falling? One answer may be found in a broad transition toward highly targeted "spearphishing" attacks, in which a much smaller number of messages are sent, each specifically tailored to legitimizing the attack to its intended target (Symantec, 2016). This familiarity lends spearphishing some of its efficacy, but the attack strategy is likely also to be bolstered by the prevalence effect.

Are hackers intentionally limiting the number of attacks they send, harnessing the power of the prevalence effect (wittingly or not) to hack the human user? The artificial depression of SP has been previously implicated as a potential dimension of cyberattack (Sawyer et al., 2016). SP of cyberattacks is something that both attackers and defenders may broadly control—the former through the ratio of bona fide attacks to "grey signals" (as in Sawyer et al., 2016); the latter through automated signal detection (spam filters, malware detection, etc.) and training resources allocated to cyberdefense. As the logarithmic line between only very small decrements and abject failure can be very fine, strategies could be formulated to push any attacked parties "over the edge." As the frequency of cyberattacks reaching users declines to very low levels, victims may become increasingly unlikely to detect and respond to remaining threats and increasingly more likely to provide hackers with compromising information. It is important to note that prevalence effects in email are not a foregone conclusion. This is because highly complex tasks sometimes exhibit minimal or nonexistent prevalence effects (cf., Adams & Humes, 1963; Lanzetta, Dember, Warm, & Berch, 1987). This is especially the case when a task is operationally diverse, as is the act of checking and responding to email, which requires first selecting, reading, and then comprehending, before formulating and inputting a response. Vigilance effects have been shown to be task-type-specific (Warm et al., 2008). Therefore, while previous efforts have indicated the presence of the prevalence effect in the complex tasks that military cyberdefenders undertake (Sawyer et al., 2014), there is a need to determine whether such prevalence effects occur in the operationally dissimilar task of checking email for cyberthreats. Further, understanding of the magnitude and pattern of any such effects will provide insight that can generalize to other tasks involving the serial inspection of messages, or indeed any sequential set of items (e.g., luggage inspection).

In cybersecurity in general, and email cybersecurity in particular, the prevalence effect may change user expectations, strategies, and behaviors in ways that make it not simply a nuisance but a potential attack vector. In a "protected" email inbox, where the SP of cyberattacks is inherently low, the prevalence effect renders the chance of failing to identify a threat to be a problematically high one. In a spam folder, an environment with a high SP of malicious emails, the prevalence effect renders the chance of failing to identify any given threat as being much lower. To complicate matters further, successful attacks are generally not immediately, or even necessarily ultimately, revealed to the person or group attacked (Sawyer et al., 2016). Even if those attacked do become aware, the rate at which successfully detected attacks are reported is unknown and, likely, actively obfuscated (see Hancock, Hancock, & Sawyer, 2015). By definition, then, the ground truth of threat prevalence is never fully specified. This stands in contrast to contexts like automotive injury, in which the result of failures to detect and respond are tragically obvious and thus are comparatively well documented. Estimates of cyberattack frequency in the real world (e.g., see Symantec, 2016) are therefore always inherently suspect—generated from algorithmic success rates that undoubtedly fail to capture the full scale, frequency, and epidemiological impact of the problem.

Experimental efforts investigating email cybersecurity are a seemingly obvious path to obtaining less biased data, but they do face their own notable challenges. Sending attack emails to personal accounts is understandably fraught

with privacy concerns. In the case of the Google breach cited previously, this includes the potential for litigation toward the involved "researcher" (*Washington Post*, 2017). While the indignation of the attacked is understandable, the fact remains that actual attackers have, comparatively, great freedom to test malware delivery strategies. On the defensive side of the equation, ethical considerations severely limit even the questions that may be asked. Consider this, for example: The robustness of Google's cybersecurity has historically led to a low probability of social engineering attacks in user inboxes, and the prevalence effect may have contributed to individuals' inability to detect and respond to such email threats. Could Google ethically remove protections from a subset of its users and observe the outcome effects? Certainly, any such revelation would result in great scrutiny of the company. Even if they could, where is the business case for such an action? Academic attempts to procure good data by recreating naturalistic cyberattack circumstances are likewise often frustrated (see Jagatic, Johnson, Jakobsson, & Menczer, 2007, for an illustrative example). However, psychophysical laboratory techniques that tightly control the timing of each trial can lack ecological validity. Email use occurs in an open-ended timeframe; users set their own pace in evaluating each new message. As such, lexical decision, vigilant attention, and other traditional laboratory tasks may each fail to capture the richness of the complex, multistage process that is the checking and evaluation of email. At their worst, such approaches may in fact result in iatrogenic effects, peripheral to the intended research questions (Hancock, 2013).

One potentially rather fruitful path forward is to employ simulation, a strategy used in many other applied settings in which risk prevents ethical experimentation in real-world settings (as in Hancock & Sheridan, 2011). The Email Testbed (ET) simulates a workplace environment and is one in which email requests are frequent, homogeneous in terms of content, and expected to be replied to promptly. This previously vetted option provides participants in the laboratory the opportunity to interact with a realistic but simulated work-email environment in the role of an administrator for a fictitious company—currently Cog Industries (cognind .com). Participants process forms containing sensitive personal information. Each incoming email must be answered by either downloading and filing a .pdf attachment or uploading an appropriate .pdf attachment. Attacks can be injected into the work schedule at any point and in any form. In a previous report, even minimal cyberdefense training has been shown to significantly boost detection of attack emails (Sawyer et al., 2015). Moreover, the success recorded by that earlier work further justifies the ET as a viable simulation tool for the present inquiry.

Therefore, in the present experiment, a simulated email environment was used to deliver cyberattack probes at three levels of SP. In order to ascertain the potential existence and form of a logarithmic decrement function, "relative accuracy" was measured in terms of the proportion of reported signals divided by total available signals in each condition. Three hypotheses were advanced. First, we hypothesized that at lower SP of email-delivered cyberattacks, participants would exhibit lower relative accuracy. Second, we hypothesized that among the three levels chosen, the pattern of relative accuracy would be better fitted by a logarithmic than a linear function (cf., Mitroff & Biggs, 2014). Third, we hypothesized that workload, as measured by the weighted global scale of the NASA TLX (Hart & Staveland, 1988), would be lower for individuals receiving email-delivered cyberattacks at a higher SP and higher for individuals receiving attacks at a lower SP (cf., Finomore et al., 2013).

METHOD

Participants

A sample of 33 participants was recruited from the undergraduate population of the University of Central Florida (UCF) and were provided extra class credit for their participation. All were required to have 20/20 or corrected to 20/20 vision and to self-report having no neurological impairments. Three were removed for failure to complete the experimental protocol, resulting in a final gender-balanced sample of 30 (n = 30). This sample size was established

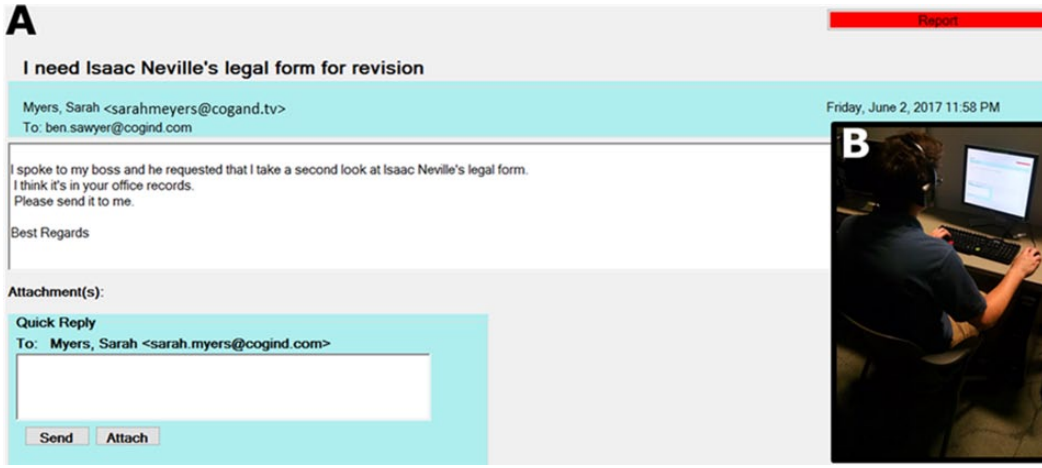

Figure 1. (A) The Email Testbed (ET) was designed to simulate interaction in common online commercial webmail interfaces. Participants received emails asking them to upload or download secure documents. Cyberattack emails had multiple cues as to their nature—in this phishing email, for example, the inbound address, ending in ".tv," and the body of the email, lacking a signature. Should a participant improperly upload and send a document, a miss would be recorded. Should the participant click on the red "Report" button, a correct detection would be recorded. Attack emails were received at rates of 1%, 5%, or 20%. (B) Each participant addressed 300 such emails. Headphones playing white noise were used to minimize distraction by ambient sound and to deliver instructions at the end.

based upon analysis of effect sizes from previous research (Sawyer et al., 2015).

Apparatus and Stimuli

Informed consent completion and the collection of demographic data were achieved via an internet-connected laptop using Qualtrics (2015). The recorded measures here were subjective workload via the NASA TLX, age, and gender. The ET (as in Sawyer et al., 2015; see Figure 1) was used to present all stimuli. Both attack and neutral emails came from an extensive corpus of validated messages (Sawyer et al., 2015) and were randomly drawn from that corpus without replacement. Participants role-played an administrative position within the fictitious Cog Industries and received emails either containing or requesting sensitive PDF attachments. After opening an email in the inbox, participants were able to (a) download attachments, (b) reply and upload their own attachments, or (c) report an email as suspicious. Upon taking one of these three actions, participants returned to the inbox, which then presented their next email. Participants therefore received and dealt with emails serially.

Legitimate emails were always delivered from addresses ending in "cogind.com" and always asked participants to either download a PDF file or upload an existing file. Attack emails were always delivered in the form of either (a) *malware* attack, as a downloadable executable (.exe) file, or (2) *phishing* attack, as a request for a form from an unauthorized outside email address ending in a .tv domain suffix. As such, attack emails contained multiple highly salient cues as to their nature. Attacks were presented at an SP of 1%, 5%, or 20%, which was a between-participants manipulation, balanced between upload and download background events. Participants were not informed as to the SP level under which they performed the task. The primary dependent measures were (a) accuracy, in terms of attack email detect-and-report rate, and (b) response time (RT).

Procedure

For each participant the task was completed in a single session. After completing the informed consent and the initial demographics, participants donned noise-canceling headphones. Each person received on-screen training with opportunities to ask questions. First, the functions

of each of the buttons in the interface were described. The participants were then instructed to report any suspicious emails, to supply PDFs to those that legitimately requested them from within Cog Industries, and to file incoming PDFs from other people within the company. A training session presented 20 example emails, 10 of which were attacks—five ".exe" malware, five phishing. Participants were required to achieve an 80% accuracy rate within the training set, and all of our samples were successful in this regard. Upon beginning the main experimental task, participants moved at a pace of their own choosing through 300 emails, taking as much time as they needed. We drew at random 1%, 5%, or 20% of those emails from a corpus of attack emails. As this arrangement resulted in an odd number of emails for the 1% and 5% conditions, alternating participants received either more malware attacks or more phishing attacks, so as to achieve a counterbalance in the total participant population. Upon the completion of all emails, participants received an audio message through their headphones as well as an email through the ET interface instructing them that the experiment was at an end. They then completed exit demographics and subjective workload scales. They were then debriefed by the research associate and left the experimentation facility. This research was approved by the Institutional Review Board at The University of Central Florida. Informed consent was obtained from each participant.

RESULTS

Data from 30 participants (n = 30) were included in the present analysis. Average time to complete the experimental task was 61.02 minutes (SD = 20.81 minutes). Prevalence effects are not themselves a product of fatigue (see Green & Swets, 1966), but we did confirm our email task itself was not unduly fatiguing by assessing the impact of total completion time upon accuracy rate, $F(1, 29) = 0.39$, $p = .94$. Resulting data were initially submitted to a mixed MANOVA to assess the impact upon accuracy and RT of 2 Attack Types (upload, download; within) × 3 SP (1%, 5%, and 20%; between) × 2 Genders (male, female). This design yielded no significant interaction or main effect of gender or attack type (the within-participant factor), and so a between-participants MANOVA was used for the analysis subsequently reported. False alarms, as in some previous investigations (Wolfe et al., 2005), occurred less than 1% of the time, and so signal detection analyses were not performed on the present dataset.

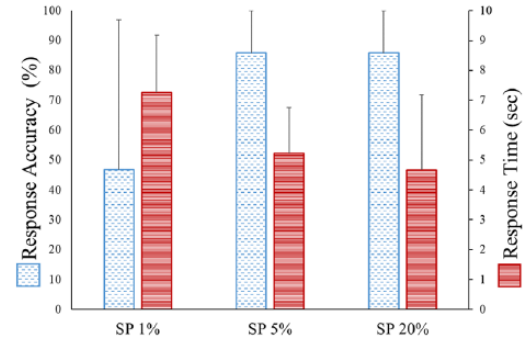

Figure 2. In email-delivered cyberattacks, a prevalence effect is seen. Despite taking more time (striped bars) in the self-paced ET task, participants in the low signal probability (SP) 1% condition detected and reported attack emails (dotted bars) at a significantly lower rate. Actual attack rates in commercial webmail systems are well below 1%, and the better the software, the lower that number. This suggests a "prevalence paradox," in which better software catches more attacks but leaves the human with an ever-smaller chance of detecting those that remain.

The present, between-participant, analysis tested the impact upon accuracy, RT, and the NASA TLX Composite Score (TLX; Hart & Staveland, 1988) of three SP levels (1%, 5%, and 20%). A significant main effect of SP was detected; Wilks' Lambda = .05, $F(6, 52) = 146.77$, $p = .02$, $\eta^2_p = .27$. Between-participants ANOVA results revealed the effect to be significant as related to both variables: response accuracy, $F(2, 27) = 4.22$, $p = .03$, $\eta^2_p = .24$, and RT, $F(2, 27) = 4.53$, $p = .02$, $\eta^2_p = .25$ (see Figure 2). These data indicated that, when the prevalence of attacks was raised, RT decreased while accuracy increased. Results additionally show that the lowest SP led to the highest levels of RT and lowest levels of accuracy. No significant effect on TLX scores was observed, $F(2, 27) = 2.08$, $p = .14$, $\eta^2_p = .13$, although nonsignificant trends (at SP 1% M = 23.88, SD = 13.21; at SP 5% M = 34.76, SD = 19.95; and at SP 20%

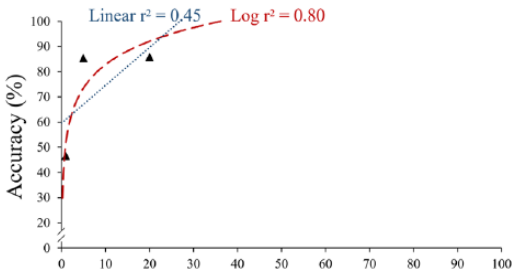

Figure 3. The logarithmic fit for these data proves superior to linear fit, suggesting that the pattern found is one of logarithmic decay of accuracy as signal probability (SP) approaches zero. This is consistent with patterns seen in past research of the prevalence effect (Mitroff & Biggs, 2014).

$M = 37.18$, $SD = 12.26$) actually show higher SPs tended to be paired with higher workload.

Finally, aggregate data from the main effects of SP on accuracy were tested for both linear and logarithmic fit. Logarithmic fit ($r^2 = .80$) proved to be superior to linear fit ($r^2 = .45$). As Figure 3 clearly illustrates, this implies that the underlying pattern of the data in fact showed one of logarithmic decay as SP approached zero.

DISCUSSION

The present data show, for the first time, the prevalence effect in email-delivered cyberattacks. In line with our first hypothesis, lower SP of email-delivered cyberattacks did result in lower relative accuracy in reporting malicious emails. In support of our second hypothesis, the pattern of response accuracy among the three SP levels chosen was indeed better fit by a logarithmic than a linear function, further confirming that the prevalence effect is at work in the present context (as in Mitroff & Biggs, 2014). The ET and accompanying email corpus provided a real-world task and attack types, providing external validity that was further reinforced by the match between our findings and data collected from actual cyberattacks (Symantec, 2016). Email-delivered cyberattacks can therefore join a rich body of other contexts in which prevalence effects and vigilance degrade human performance, including medical imaging, air traffic control, friendly fire incidents, and IED detection, to name but a few. Further, our findings suggest that the sizable literature on both vigilance and the prevalence effect can be brought to bear successfully upon this emerging and crucial domain.

Contrary to our third hypothesis, subjective workload showed no significant effect of SP level. Nonsignificant trends were in the direction of greater workload for greater SP conditions. It is worth noting that the present study, in common with a number of other studies investigating prevalence effects, was a self-paced task such that movement rate through the task was determined by the participants themselves. The lack of pattern regarding workload may suggest that when SP manipulations are dissociated from the rate at which events are presented, workload itself becomes dissociated from performance (Hancock, 1996, 2017). Note that, at lower levels of SP, variance increases markedly even as performance declines. This pattern would render self-assessment of performance nearly impossible, at the individual metacognitive as well as the team performance levels. It is possible the workload profile of email in general, or our task specifically, may be quite muted (and see Greenlee et al., 2016; Young, Brookhuis, Wickens, & Hancock, 2015). Of course, our email-delivered cyberattacks were quite salient, and the email corpus used was presented serially, a pattern of email communication common in clerical workplace settings where filing and providing documents may represent the main function of an individual worker. This pattern differs markedly from email workflow in academic, managerial, or many other workplace settings where email is more tertiary, strategic prioritization is more necessary, and uninterrupted stretches of 300 emails are hopefully rare. In a task that provided less salient signals, constrained event rate, or allowed participants to engage with emails in the order of their choosing, different results would be expected in terms of workload and, indeed, potentially in terms of all study dependent variables. Further research is certainly needed.

The broad idea substantiated here, that in human-machine teaming, cyberattacks may target human cognitive vulnerability rather than algorithmic shortfalls, has significant implications. To use the analogy of poker, the field of cybersecurity has been playing the cards, but the

possibility now exists to directly play the players. To be clear, this vector of attack extends beyond social engineering (as in Anderson, 2008) to exploitation of fundamental shortcomings of human information processing. Indeed, Symantec (2016) provides circumstantial evidence that attackers may already be strategically reducing SP, and thereby enjoying enhanced cyberattack efficacy (Ponemon Institute, 2016). While these parties may not be aware of the cognitive underpinnings of their success, ethically unrestrained A-B testing on the large sample sizes would readily reveal this pattern. Other potential vectors of human-mediated digital attack certainly exist; the basic and applied psychological literature is rich with effects revealing the limitations of human cognition. Just as the visual perception community has painstakingly catalogued hundreds of visual illusions and the advertising industry catalogued manipulations of decision-making, so now must cyberdefense embark upon understanding which features of human cognition may be repurposed as attacks or to buoy defense. The solution to such vulnerabilities would, at first glance, seem to be more rigorous algorithmic protection.

In considering protections for human operators, the authors found a paradoxical quandary: The SP manipulation in the present experiment can be as easily couched in automated protection as it could be in cyberattack strategy. Success on the part of algorithmic defenses would directly engender less successful, prevalence-effect influenced human partners. We have coined the term "prevalence paradox" to describe the counterintuitive situation under which increasingly effective protective autonomy leads to increasingly few critical signals and so an ever-growing likelihood that human operators will miss the signals that do remain. To be clear, this in no way means the user should disable their algorithmic protections. In absolute terms, such machine detection enhances overall performance of the human-machine system. Still, in any teaming situation, success of one teammate at the expense of another is an obvious problem, fraught with questions of blame. In the present environment of email cyberattack, the penetration of even a single attack email into a user's inbox can be seen by that user as a direct indictment of the efficacy of protective systems (spam filters, etc.), regardless of the absolute reality of the situation. A brief thought experiment reveals that the prevalence paradox is therefore not only a performance issue but also tightly coupled with the emergent property of trust (Hancock et al., 2011). Prevalence paradox effects may, for example, be strong arbiters of usage, as levels of trust determine levels of engagement spanning from overreliance to neglect (Lee & See, 2004). While there is presently no work that explicitly explores this link among prevalence effects, teaming with automation, and trust, there is evidence that such a relationship does exist. Work linking automation to trust (e.g., Molloy & Parasuraman, 1996; Parasuraman & Riley, 1997) suggests that monitoring of and by autonomous systems can lead to miscalibration of trust and resultant suboptimal performance or even abandonment altogether of a system that might otherwise provide benefit. Further, previous literature clearly shows that users who distrust or overtrust tools are subject to a range of vulnerabilities (Parasuraman & Riley, 1997), including mistrust and miscalibrated trust. Prevalence paradoxes therefore offer foundational points upon which to build deception and exert influence, sabotaging trust relationships within the human-machine teaming that is so crucial to cyberdefense.

It is also important to consider the fact that the prevalence paradox certainly exists in contexts beyond cyberdefense. More generally operationalized, this construct describes any situation in which an attempt to improve performance by the interception of threats has the unintended effect of reducing the probability of signals to the point of inducing a prevalence effect. Therefore, the construct's impact will very likely extend to contexts beyond cybersecurity, and even human-machine integration, to affect even human-human teaming. Consider how highly reliable individuals may unintentionally intercept enough signals within a context to expose their human teammates to prevalence effects. In medical settings, certain types of common medical or iatrogenic issues might be intercepted with great efficiency by skilled nursing staff. Physicians, so protected from the base rate of these issues, might be at a great

disadvantage in their own detection. Pilots of commercial aircraft might be guided by autonomous systems successfully minimizing certain error conditions. The base rate of such incidents may then drop below the ability of pilots to detect and respond. Finally, consider drivers protected by lane-centering technologies that apply small corrections to steering to keep the vehicle centered in the lane. Users might well find that drift from the lane becomes so rare that they are unable to detect exceptions in a timely manner. Teaming in these examples involves humans and machines, or humans and humans. The common thread is one of robust assistance that reduces but does not completely eliminate threats and leads to later prevalence effects for human stakeholders. Excellent medical staff are not infallible, nor is aircraft automation, nor is ground vehicle automation, but the closer they come to perfection, the more likely they are to elicit a failure downstream. Trust is at risk in every context we reference here, as it is in the present cybersecurity context. Physicians' trust in their staff, pilots' trust in their autopilot, and drivers' trust in their cars' advanced safety technology all directly impact their role in teaming. Here, distrust might cause individuals to discontinue use of a resource that would otherwise protect them. The prevalence paradox, therefore, may point to a more universal quandary: How is it possible to provide excellent protection while still allowing those you protect an unhindered opportunity to protect themselves?

What design interventions might defuse the prevalence paradox? While the present results provide no explicit solution, exploring previous inquiries reveals some paths forward. This issue, its consequences in large-scale cyberdefense, and possibilities for so-called "prevalence attacks" have been previously discussed (Sawyer et al., 2015). The present work is experimentally grounded in the psychological phenomenon of vigilance and the associated vigilance decrement (Warm & Jerison, 1984), of which the prevalence effect, referenced here, is one element. In repetitive observations in search of rare signals with little control over the occurrence of the next target, it is common to see a decrement in detection rate. Such vigilance decrements are strongly affected by SP via "event rate," the speed at which new candidate stimuli are presented. Experimental efforts evaluating tasks of this nature frequently forego an analysis of overall rate of events and rather focus upon the RT of participants to individual events (as in Wolfe et al., 2005, 2007). This, in part, is because such RTs provide a window into participants' strategies. For example, speed/accuracy tradeoffs can exist in which lower accuracy is associated with moving through material at a faster pace (as in Fleck & Mitroff, 2007). The question as to whether moving through trials too quickly is a basis for the prevalence effect has been asked, and encouraging longer consideration before rendering a decision does have an impact in terms of reducing missed signals (Wolfe et al., 2007). From the design of user experience (UX) standpoint, enforcing a certain time for scrutinizing an incoming email is unlikely to be embraced by end-users, but artful renderings might find an audience. Of course, just as such microstrategies may have efficacy, event rate may potentially be influenced by altering the macrostructure of demand in the workplace (see Sawyer et al., 2016). The design of methods to help better calibrate overall staff levels to aggregate threat levels over time may thus be a fruitful path forward.

Event rate is not the only potential point of intervention. In considering the levels of SP chosen for the present experiment, it became obvious that further exploration at lower SP levels may reveal more granular patterns and so expose novel countermeasures. It is also possible the different types of attacks have unique signatures of performance decay, and efforts to understand these patterns could lead to better strategies in cyberdefense. For example, if some forms of phishing attacks become difficult to respond to at higher SP than others, these could be more aggressively filtered, trading higher false positives for a better detection rate. SP inflation through the injection of "pseudo-signals," attack emails that deliver payloads for training, is a strategy already being attempted as a cyberdefense strategy. Such efforts and products seem to focus on training alone without an apparent understanding of the cognitive underpinnings described here. Attempts of this nature, for example, the random intermittent injection of

warning messages, run afoul of the exceptional ability of humans to categorize the threats and nonthreats in their environment. Finding a "training email" tells a user nothing about a real threat in the environment, and as such that user is unlikely to change his or her criterion. Sending real threats, of course, is a very problematic strategy. One possible, although untested, middle ground would be "defanging" of existing attacks, by stripping payloads and links to attackers. These could be as simplistic as messages regarding thwarted attacks, but such defanged attacks might also be reserved and reinjected, even to other users of the system. In this role, defanged injections might serve a more ecologically valid role, representing the true SP of attacks in the email ecosystem and suggesting better calibrated digital hygiene (Sawyer et al., 2015). Indeed, the line here is to find one, and once users begin thinking of these attacks as "not real," their efficacy would dissipate.

It is important to recall the unique advantages for defense of the entirely synthetic environment of cyber. The strategies we discuss might prove impractical in settings such as radiology, where there is real risk of overlaying a fake signal on real information, such as the shadow of an undetected tumor (see Wolfe et al., 2007). The serial nature of email presentation largely eliminates the latter concern but does not eliminate the problem that such false flag attacks on the part of the email system might themselves have serious effects on trust or user understanding of the probability of actual attacks. Likewise, a strategy suggested by Wolfe and colleagues (2007) in which observers might be "retrained" by exposure to high SP epochs of search activity might have real utility in email. Negative effects may be magnified during rapid digital "movement" between high- and low-SP environments, as humans require time to adapt to changed levels of SP (Wolfe et al., 2007). Design to encourage advantageous transitions is worth investigation. The "defanged" emails collected by automated systems might be more strategically deployed under such a strategy. Moreover, such a cache of high-SP targets already exists in spam folders, and so by enforcing spam reviews on a regular schedule, some benefit might be gained. Finally, state detection solutions may allow for greater transparency and communication between human and machine team members. Research monitoring visual behavior and physical interaction with systems in tandem has progressed greatly in recent years (see Lee et al., 2017, for a surface transportation example with potential application for other interfaces) and might allow state detection solutions in many varieties of interface, including those in cyberdefense. Likewise, electroencephalographic evoked response potentials (ERPs) for error detection in combination with the email injection strategies noted above might shed light on operators' ability to detect errors or signals (as in Sawyer, Karwowski, Xanthopoulos, & Hancock, 2017), then allow for immediate modulation of the cyber environment. Extant inquiry contains many more potential ameliorative strategies, each of which must be redesigned and evaluated in the specific context of email cyberdefense. Significant rewards await the individual, commercial entity, or nation-state finding workable solutions.

It might seem prudent to remove operators entirely from the dangers of email cyberattack, perhaps through the development of putatively "perfect" algorithmic cyberdefense. For advocates of removing the human from the loop of control, a number of cautions should be emphasized. First, it is important to remember that human cognition is, as of this writing, the superior general-purpose signal detector in most human-machine team contexts. While machines are widely acknowledged to be superior at certain types of problems (see De Winter & Hancock, 2015), humans have their own strengths. Indeed, humans and machine learning may encounter different difficulties in the same search. Situations where both machines and humans struggle with the complications of the prevalence effect may be well represented through signal detection theory (SDT) tradeoffs (see Warm & Jerison, 1984). This balance between false alarms and misses has not been *perfectly* addressed in human biology, as we have shown, nor in any known biology (see Bond & Kamil, 2002, for examples in other organisms), but through the prolonged actions of evolution, it may be close to optimally addressed. Any imagined "perfect" algorithmic cyberdefense system

would need to restrike this balance in a way not only superseding these existing biological solutions but in agreement with their edge cases. Indeed, in situations where prevalence and vigilance effects are well-controlled, the flexible cognition of humans is well-suited to monitor the unpredictable edge cases of autonomy.

Consider, for example, the issue of machine false-positives that result in legitimate data caught in the digital dragnet of a spam folder or less user-accessible location. Here, it is important to remember a second point: that human operators are fully willing to deceive and disable automation in order to achieve their own goals. Wholesale blocking of archive files by many email systems, a well-intentioned removal of the human from the loop in a cybersecurity context, has resulted in diverse renaming schemes (in which the suffix of a well-understood archive type is replaced; i.e., .zip to .piz). This workaround nicely solves the user problem, at the cost of complicating security in a fashion that undermines algorithmic cybersecurity and ultimately works to the advantage of attackers. In general, human users have the cognitive ability and environmental flexibility to become the final arbiter of the legitimacy of any digital message. There is good reason to give them such autonomy: Human signal detection is often generalized enough to identify the failures of machine signal detection, while the reverse is seldom true (Parasuraman, Masalonis, & Hancock, 2000). This asymmetry necessitates teaming, in which humans and machines alike assist with the shortcomings of the other, a state that must arise from careful design. Before demonizing operators that turn against their digital systems, it is important to remember that such behavior is a sign of a breakdown of necessary human-machine teaming.

Perhaps most fundamental, however, is not to degenerate into narratives concerning weak humans, malevolent machines, blame, and retribution. The intent, after all, is for user and machine to be on the same team. What is vitally needed now is investment in human-centered understanding of human-machine symbiosis toward actionable improvements. Cybersecurity is an exemplary case of such human-automation teaming. Despite this truth, cyberattack and concomitant cyberdefense is presently heavily, and in many cases even exclusively, vested in algorithmic and software realms. Yet here, we protest that these issues involve at least a major, if not a majority concern, with the human dimension. In this work, we have raised and demonstrated a prevalence effect in this domain and further identified a complication of this effect, the prevalence paradox. We believe there to be strong and productive links between these constructs and the traditional human factors realm of vigilance and sustained attention (Hancock, 2013) as well as other elements of human factors science (e.g., perception, decision-making, automation-mediation, etc.). The time to adapt and apply this plentitude of hard-won understanding is now, as the status quo of largely algorithmic defense investment is failing. Forces of cyberattack presently outstrip forces of cyberdefense (Gutzwiller, Fugate, Sawyer, & Hancock, 2015). Breaches of increasingly prodigious proportion (Ponemon Institute, 2016; Symantec, 2016) are becoming daily events with mounting impact. Human knowledge, disproportionately in the hands of the few, is now changing hands in quantities that must surely represent the greatest illicit exchange of information in human history. A defensive solution, we submit, can only come from expanded and continued investment in understanding the vital human factor of cybersecurity.

ACKNOWLEDGMENTS

Data reported in the paper were collected and are archived at The University of Central Florida's MIT^2 Laboratory. For further information, please contact the corresponding author. We thank the U.S. Air Force's 711th Human Performance Wing, specifically Drs. Benjamin Knott, Victor Finomore, and Gregory Funke and Mr. Brent Miller and Mr. Allen Dukes, for support in the design and creation of the ET. We thank the entire Sawyer Team at the MIT^2 Laboratory for assistance in creation of the email corpus as well as data collection. We thank Ms. Szuhui Wu for assistance with Data Science, especially consolidation and assurance. We thank Drs. Mouloua, Szalma, Matthews, and Warm for constructive criticism and feedback. The latter's unparalleled knowledge of the foundational concepts within this work, delivered during the early careers of both authors, was invaluable. Thank you, Joel.

KEY POINTS

- The prevalence effect, here demonstrated for the first time in email cybersecurity, refers to rare signals in an environment being substantially more difficult to detect, even taking into account their low occurrence.
- Cyberattackers may be intentionally inflicting prevalence effects by delivering attacks at a lower per-user rate. Such hacking of the human may explain why email phishing attacks have declined in number while growing in impact.
- Under the prevalence paradox, as helpful automation reduces the number of attacks, human operators are increasingly likely to fail to detect and report remaining attacks.
- The prevalence paradox has strong implications for human-machine teaming and trust.
- We suggest that efforts to remove the human from the loop are likely ill-fated and instead suggest design interventions to mitigate prevalence effects and strengthen human-machine teaming.

ORCID ID

Ben D. Sawyer https://orcid.org/0000-0002-3009-6707

REFERENCES

Adams, J. A., & Humes, J. M. (1963). Monitoring of complex visual displays: Training for vigilance. *Human Factors*, *5*(2), 147–153.

Anderson, R. (2008). *Security engineering*. New York, NY: John Wiley & Sons.

Bond, A. B., & Kamil, A. C. (2002). Visual predators select for crypticity and polymorphism in virtual prey. *Nature*, *415*(6872), 609–613.

De Winter, J. C. F., & Hancock, P. A. (2015). Reflections on the 1951 Fitts list: Do humans believe now that machines surpass them? *Procedia Manufacturing*, *3*, 5334 –5341.

Finomore, V. S., Shaw, T. H., Warm, J. S., Matthews, G., & Boles, D. B. (2013). Viewing the workload of vigilance through the lenses of the NASA-TLX and the MRQ. *Human Factors*, *55*(6), 1044–1063.

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science*, *18*(11), 943–947.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Greenlee, E. T., Funke, G. J., Warm, J. S., Sawyer, B. D., Finomore, V. S., Mancuso, V. F., & Matthews, G. (2016). Stress and workload profiles of network analysis: Not all tasks are created equal. *Advances in Human Factors in Cybersecurity*, 153–166. Springer, Cham.

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

Hancock, P. A. (1996). Effects of control order, augmented feedback, input device and practice on tracking performance and perceived workload. *Ergonomics*, *39*(9), 1146–1162.

Hancock, P. A. (1997, April). *Hours of boredom, moments of terror, or months of monotony, milliseconds of mayhem*. Paper presented at the Ninth International Symposium on Aviation Psychology, Columbus, OH.

Hancock, P. A. (2013). In search of vigilance: The problem of iatrogenically created psychological phenomena. *American Psychologist*, *68*(2), 97–103.

Hancock, P. A. (2015). *Hoax springs eternal: The psychology of deception*. Cambridge, UK: Cambridge University Press.

Hancock, P. A. (2017). Whither workload? Mapping a path for its future development. In L. Longo & M. C. Leva (Eds.), *H-Workload* (pp. 1–15). New York, NY: Springer.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

Hancock, P. A., Hancock, G., & Sawyer, B. D. (2015). Cybernomics and the implications of cyberdeception. *The Ergonomist*, *537*, 12–14.

Hancock, P. A., & Sheridan, T. B. (2011). The future of driving simulation. In D. L. Fisher, M. Rizzo, J. K. Caird, & J. D. Lee (Eds.), *Handbook of driving simulation for engineering, medicine and psychology* (pp. 1–11). Boca Raton, FL: CRC Press.

Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress sustained attention. *Human Factors*, *31*, 519–537.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: Elsevier.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM (Association for Computing Machinery)*, *50*(10), 94–100.

Lanzetta, T. M., Dember, W. N., Warm, J. S., & Berch, D. B. (1987). Effects of task type and stimulus heterogeneity on the event rate function in sustained attention. *Human Factors*, *29*(6), 625–633.

Lee, J. B., Sawyer, B. D., Mehler, B., Angell, L., Seppelt, B., Seaman, S., Fridman, L., & Reimer, B. (2017). Linking the detection response task and the AttenD algorithm through the assessment of human machine interface workload. *Transportation Research Record*, No. 17-06664.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Maybury, M. T. (2012). *Cyber Vision 2025: United States Air Force Cyberspace Science and Technology Vision 2012-2025* (SAF/PA Public Release Case No. 2012-0439/460/715). Washington, DC: US Air Force.

Mitroff, S. R., & Biggs, A. T. (2014). The ultra-rare-item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science*, *25*(1), 284–289.

Moar, J. (2017). *The future of cybercrime & security*. Basingstoke, UK: Juniper Research.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors*, *38*(2), 311–322.

Parasuraman, R., Masalonis, A. J., & Hancock, P. A. (2000). Fuzzy signal detection theory: Basic postulates and formulas for analyzing human and machine performance. *Human Factors*, *42*(4), 636–659.

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

Ponemon Institute. (2016). *2016 cost of cyber crime study & the risk of business innovation*. Traverse City, MI: Author.

Qualtrics. (2015). *Online survey generator* [computer software]. Provo, Utah: Author.

Sawyer, B. D., Finomore, V. S., Funke, G., Mancuso, V., Warm, J. S., & Hancock, P. A. (2015). Evaluating cybersecurity vulnerabilities with the email test-bed: Effects of training. *Proceedings of the 19th Triennial Congress of the International Ergonomics Association*, *9*, 14.

Sawyer, B. D., Finomore, V. S., Funke, G., & Warm, J. S. (2014). Cyber vigilance: Effects of signal probability and event rate. *Proceedings of the 2014 Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 1771–1775.

Sawyer, B. D., Finomore, V. S., Funke, G., Warm, J. S., Matthews, G., & Hancock, P. A. (2016). Cyber vigilance: The human factor. *American Intelligence Journal*, *32*(2), 157–165.

Sawyer, B. D., Karwowski, W., Xanthopoulos, P., & Hancock, P. A. (2017). Detection of error-related negativity in complex visual stimuli: A new neuroergonomic arrow in the practitioners' quiver. *Ergonomics*, *60*(2), 234–240.

Schultz, N. B., Matthews, G., Warm, J. S., & Washburn, D. A. (2009). A transcranial Doppler sonography study of shoot/don't-shoot responding. *Behavior Research Methods*, *41*(3), 593–597.

Symantec. (2016). *Internet Security Threat Report Volume 22*. Mountain View, CA: Author.

Szalma, J., Schmidt, T., Teo, G., & Hancock, P. A. (2014). Vigilance on the move: Video game-based measurement of sustained attention. *Ergonomics*, *57*(9), 1315–1336.

Warm, J. S., & Jerison, H. J. (1984). The psychophysics of vigilance. In J. S. Warm (Eds.), *Sustained attention in human performance* (pp. 15–60). Chichester, UK: Wiley.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*(3), 433–441.

Washington Post. (2017). Why this Google Docs phishing attack is particularly sneaky. Retrieved from https://www.washingtonpost.com/news/the-switch/wp/2017/05/03/why-this-google-docs-phishing-attack-is-particularly-sneaky/

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: Rare items often missed in visual searches. *Nature*, *435*(7041), 439–440.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*, *136*(4), 623–638.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17.

Ben D. Sawyer, postdoctoral associate at the Massachusetts Institute of Technology's AgeLab and Center for Transportation and Logistics, received both a PhD in Human Factors Psychology in 2015 and an MS in Industrial Engineering in 2014 from the University of Central Florida. For more information and press coverage of his work, visit bendsawyer.com..

Peter A. Hancock, Provost Distinguished Research Professor, Pegasus Professor, and University Trustee Chair at the University of Central Florida's Department of Psychology, also holds appointments with The Institute for Simulation and Training (IST), Industrial & Systems Engineering, & Civil Engineering. He received a PhD from the University of Illinois in 1983, and a D Sc in Human-Machine Systems from Loughborough University in 2001.

Date received: June 3, 2017
Date accepted: May 3, 2018