Association for Information Systems
AIS Electronic Library (AISeL)

SIGHCI 2018 Proceedings

Special Interest Group on Human-Computer Interaction

12-13-2018

Creating a Realistic Experimental Scenario for HCI-Based Deception Detection Research with Ground Truth and Unsanctioned Malicious Acts

Michael D. Byrd
University of Arizona, byrd@email.arizona.edu

Jeffrey L. Jenkins
MIS, University of Arizona, Tucson, AZ, United States., jjenkins@byu.edu

Joe S. Valacich
University of Arizona, valacich@email.arizona.edu

Parker A. Williams
University of Arizona, parkerwilliams@email.arizona.edu

Creating a Realistic Experimental Scenario for HCI-Based Deception Detection Research with Ground Truth and Un-Sanctioned Malicious Acts

Michael D. Byrd
The University of Arizona
byrd@email.arizona.edu

Jeffrey L. Jenkins
Brigham Young University
jjenkins@byu.edu

Joseph S. Valacich
The University of Arizona
valacich@email.arizona.edu

Parker A. Williams
The University of Arizona
parkerwilliams@email.arizona.edu

ABSTRACT

This research-in-progress note reports on the design and execution of a study in HCI-based deception detection. The objective of the study is to examine the impact of knowledge of tracking and countermeasures on the neuro-motor changes detected when subjects commit a malfeasant act. To examine this, an experimental context and design was required that would afford the subjects an opportunity to commit an un-sanctioned malicious act while tracking ground truth in an unobtrusive manner. The experimental design, study execution, and preliminary results are presented.

Keywords

Psycho-Physiological Deception Detection; PDD; Countermeasures; Attentional Control Theory; ACT; Ground Truth; Unsanctioned Lie; Deception Experiment; HCI-based Deception Detection

INTRODUCTION[1]

Deception and malfeasance are endemic in our world today. Crime and terror threats are never far from top of mind and the enhanced security measures put in place in reaction impact us all on almost a daily basis. Above and beyond questions of law enforcement, safety, and security, the economic costs of malfeasance can be staggering. The Association of Certified Fraud Examiners estimates that in 2018 the typical organization loses 5% of their revenues to fraud (ACFE 2018) and the insurance industry alone is estimated to have lost over $80 billion in 2015 (Coalition Against Insurance Fraud 2015). Individual instances can be much more spectacular and expensive (e.g., Enron, WorldCom) and can systematically infiltrate and undermine entire organizations such as Wells Fargo with their fake accounts scandal (Conti-Brown 2017). In addition, insider threats have become one of the most critical threats facing government agencies and corporations (Gorman 2014).

To address these threats, new approaches to screening are needed. As information technology has revolutionized the conduct of business operations, it is now poised to dramatically alter security and screening processes. Specifically, research into the use of information systems for deception detection has grown substantially in scope in recent years (Nunamaker Jr et al. 2016). Systems that utilize technologies such as eye tracking (Proudfoot et al. 2016), facial recognition (Su and Levine 2016), and linguistic analysis (Burgoon et al. 2016) are increasingly being tested and deployed from the lab into real world usage. Previous work has indicated that such deviant behaviors can also be detected using commonly available human-computer interaction (HCI) devices such as computer mice (Hibbeln et al. 2014, 2017). When an individual engages in a deviant behavior there are multiple impacts on their cognitive processes. These impacts manifest as changes in the motor nervous system. These changes are detectable by analyzing movement information collected from HCI devices. The use of commonly available HCI devices offers options for deployment of such behavioral monitoring systems at scale. Expensive eye-tracking hardware or special cameras are not required – merely software that captures and analyzes the appropriate data from the HCI device. Systems that flag such HCI behaviors are being commercialized and evaluated for use in multiple contexts. These systems work by generating a baseline of a user's movements while they are engaged in innocuous activity and comparing that baseline to their movements when they answer questions relevant to the risky behavior. These systems are being deployed in

[1] This work was supported in part by the National Science Foundation under Grant No. SES-1615696 (SaTC). The authors disclose an outside interest in Neuro-ID, Inc. Conflicts resulting from this interest are being managed by The University of Arizona in accordance with its policies.

commercial contexts, such as loan underwriting, and national security contexts, such as insider threat detection.

When deploying such a system in practice, a common concern that arises is that of countermeasures. We define countermeasures in this context to mean the use of movement techniques designed to defeat the ability of a mouse movement tracking system to infer valid results. In other words, what happens when users know that they are being monitored and attempt to defeat the system? To address this question, we conducted an experiment in which we gave subjects the opportunity to choose to perform a malfeasant act (i.e., cheating) in a context in which we could know ground truth without overly intrusive monitoring. This preliminary report focuses on the context and experimental design to achieve these objectives.

BACKGROUND

This research is an extension of previous work in HCI-based detection of cognitive states and deception (Hibbeln et al. 2017; Jenkins et al. 2019). As in these previous studies, we leverage Attentional Control Theory (ACT) and the Response Activation Model (RAM) to connect these cognitive and emotional effects to motor-nervous system phenomena that are measurable using HCI devices. ACT states that experiencing negative emotions leads one's attention to shift from being goal directed to being stimulus driven (Eysenck et al. 2007). As a result, additional cognitive resources are consumed, leading to a degradation of motor performance (see Hibbeln et al. 2017). RAM states that as possible movement choices are entered into working memory, the motor-nervous system pre-plans destinations and sub-movements. This pre-planning shows up as increased changes to the actual trajectories taken in response to questions about malicious acts (i.e., less movement accuracy). Thus, we would expect to see similar changes to motor nervous system metrics as in the previous work (e.g., greater deviation and distance for cheaters than for non-cheaters). For this study, our research questions was how these effects change as additional information is provided to subjects about the fact that they are being tracked, the nature of the tracking, and potential countermeasures to the system.

METHOD

The system we used to track movements is derived from a commercial system. A small JavaScript is embedded in web pages containing the questions to which the user responds. Our system collects the same raw data as the commercial system – XY coordinates and timestamps. The raw data is then uploaded to a cloud-hosted analysis engine. The commercial system uses the raw data to generate an extensive set of features, analyzes them, and then generates results using proprietary algorithms. For this analysis, we restrict the features to a limited set of indicators (i.e., maximum deviation, normalized area-under-the-curve, and normalized additional distance).

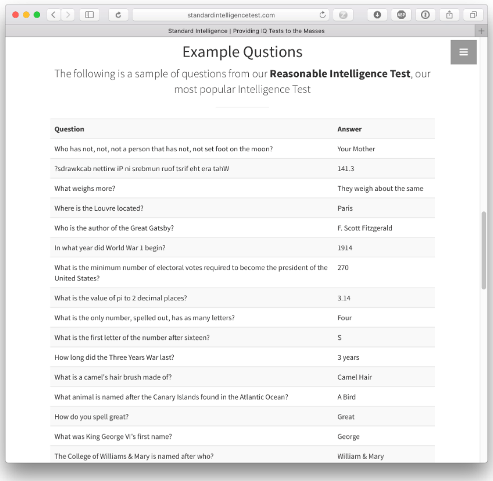
Task

Next, a task was needed in which the user could decide to commit a deviant act. It is important that the user be able to choose whether or not to commit the act. Since the detection system relies on cognitive and emotional responses to generate the changes in the motor nervous system, it is crucial that the deviant act not be sanctioned by the experimenters. If the act is sanctioned, the subjects will not have the underlying psychological and physiological responses to having performed something deviant – they will "know" that what they did was authorized. This requirement makes it a challenge to design an experimental task. The subject must be left free to decide whether or not to perform the deviant act. This also increases required sample sizes since subjects who do not decide to perform the deviant act above and beyond those needed for a statistically valid comparison group are effectively wasted.

To encourage participants to perform the deviant act, the nature of the task needs to be something they feel they can "get away with." By obfuscating the participants' activity, they are given a sense of freedom to perform an act that they might otherwise not perform due to social or legal pressure. The realities of performing experimental research using human subjects restricts the types of deviant acts available. We needed one that would be sufficiently frowned upon to evoke the desired cognitive and emotional responses, yet not so far gone as to do lasting psychological damage from having committed it. We chose to create a simulated intelligence test. This test consisted of multiple choice and fill-in-the-blank questions. This instrument is appropriate as there are many of these types of tests available online and the outcomes are of interest to participants without directly leading to potential tangible harms.

To properly analyze the outcomes of the experiment sessions, we must know the ground truth – did the participant cheat or not? If the decision as to whether to perform the deviant act is left up to the subject and the task is designed to allow suitable expectation that they will not be immediately caught and sanctioned for such performance another challenge arises – knowing if they actually performed the act. If it is obvious to the participant that their actions will be detected, they may not decide to perform the deviant act. We could have attempted to create additional software that would detect if a user visited a different web page. Alternatively, we could have attempted to generate cookies or another tracking mechanism that would have allowed us to correlate visits to our cheating websites. Neither of these approaches were satisfactory given the variety of mechanisms a user could have used to cheat. For example, a user filling out the intelligence test on a desktop or laptop could have searched for the answers using their smartphone. In other words, attempting to track the process or mechanism by which they cheated would be a losing battle. In order to be as certain as possible that the user cheated, we needed the actual outcome of the interaction to be useful to determine if they cheated. To do this, we created the fake intelligence test in such a way that the only way to achieve a top score would be to cheat.

To accomplish this, we designed the test with easy, hard, and impossible questions. Easy questions are questions that anyone should be able to answer; e.g., what is 2+2? Hard questions are questions that it is not likely that someone would know, but ultimately possible to know; e.g., obscure facts. Impossible questions are questions that it is not possible for someone to know the answer unless they cheat; e.g., made up facts (e.g., "Dr. Jeffrey Zaverik is noted for the discover of which of the following?"). We then created websites containing the answers to all types of questions and posted them online. We performed Search Engine Optimization (SEO)[2] to help insure the sites would be easy to find with a Google or Bing search. Two websites were created: one was designed to look like the vendor of the test had placed sample questions online (Figure 1); the second was designed to look like a third party was sharing the answers. These two types of sites were selected as they exhibit face validity – i.e., it is reasonable that they would exist and have the answers posted.


standardintelligencetest.com
Standard Intelligence | Providing IQ Tests to the Masses
ABP
+
≡

Example Qustions

The following is a sample of questions from our **Reasonable Intelligence Test**, our most popular Intelligence Test

Question	Answer
Who has not, not, not a person that has not, not set foot on the moon?	Your Mother
?sdrawkcab nettirw iP ni srebmun ruof tsrif eht era tahW	141.3
What weighs more?	They weigh about the same
Where is the Louvre located?	Paris
Who is the author of the Great Gatsby?	F. Scott Fitzgerald
In what year did World War 1 begin?	1914
What is the minimum number of electoral votes required to become the president of the United States?	270
What is the value of pi to 2 decimal places?	3.14
What is the only number, spelled out, has as many letters?	Four
What is the first letter of the number after sixteen?	S
How long did the Three Years War last?	3 years
What is a camel's hair brush made of?	Camel Hair
What animal is named after the Canary Islands found in the Atlantic Ocean?	A Bird
How do you spell great?	Great
What was King George VI's first name?	George
The College of Williams & Mary is named after who?	William & Mary

Figure 1: Cheating Website

Since there was no way a participant could know the answers to the impossible questions, if they got them correct they either guessed or cheated. We placed five impossible questions on the test, four multiple-choice and one fill in the blank. Each multiple-choice question had five possible answers, thus the odds of guessing an individual question are 1/5 = 20%. The fill-in-the-blank question asked for a single letter (The most used letter in the Nekmunnit language is translated to which letter in English?), thus the odds of guessing correctly are 1/26 = 3.85%. The combined probability of getting all five correct by guessing is 1/16,250. For a participant to guess three of the four multiple choice questions and the fill-in-the-blank question correctly, the odds are 1/3,250. We anticipated using on the order of 750 participants (50 per cell * 5 conditions / 33% estimated cheat rate = 750), thus we felt it appropriate to deem anyone having answered three or four of the multiple choice and the single fill-in-the-blank question correct as having cheated. After further pilot testing we dropped the hard category since it did not contribute anything to the cheating calculation and just made the test longer. We also stopped randomly selecting questions from a pool. We did this in order to more easily be able to score the questionnaires in real time to assign subject payments. In addition, we judged the additional control we gained from being able to tweak the questions outweighed any risk of order effects.

Measurement

Following the fake IQ test questions, we first asked the participants a series of demographic questions (such as gender and world region). These questions serve to collect demographic information to use in the subsequent baseline questions. We then asked the participants to confirm their demographic information with a series of Yes/No questions. These questions serve to calculate the baseline movement information as there is no reason to expect that someone would be deceptive about confirming information that they just provided. Next, we asked the participants the risk relevant questions (i.e., "I cheated on the intelligence test", "I looked up answers while completing the test"). We asked about cheating in multiple ways since our early pilots indicated that different individuals may have different interpretations of what it means to cheat.

Manipulation

In order to instantiate countermeasures in our context we gave participants varying degrees of awareness of the tracking. This manipulation was inserted immediately before the participants were asked the demographic questions described above. The control group (Group 0) was given no knowledge that their mouse movements were being tracked. The first condition (Group 1) was told simply that they were being tracked:

[2] The SEO process led us to iteratively revise our questions. For example, we originally had a question: "Dr. Jeffery Smith is noted for the discovery of which of the following?" We ultimately changed "Smith" to the made up and unique name "Zaverik" such that it could become a unique keyword for the search engine and be more likely to return our sites.

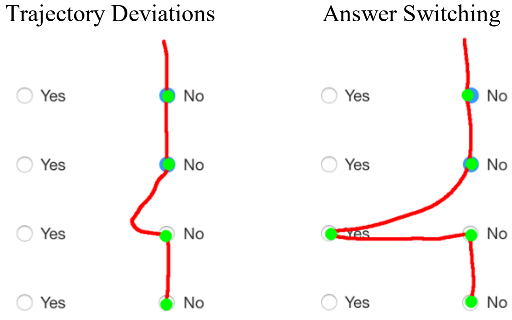
Please note: We capture your interaction data (e.g., mouse movements) and use this information in order to identify those who are being deceptive when answering these questions.

The second condition (Group 2) was given additional details about what we were looking at:

Interaction data attributes that we monitor include:

- *Behavioral indicators such as hesitation, switching answers, skipping questions and returning, etc…*
- *Movement indicators such as click accuracy, deviations from ideal trajectories, number of clicks, hesitations, etc…*

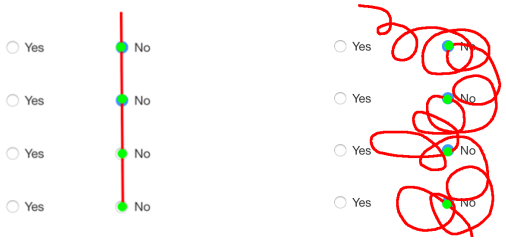
The third condition (Group 3) was shown all of the above plus was shown images of what the monitored attributes look like (e.g., Figure 3).


Trajectory Deviations

Answer Switching

Figure 3: Group 3 Manipulation (example)

The fourth condition (Group 4) was shown all of the above, plus was shown images of what some potential countermeasure strategies might look like (e.g., Figure 4).



Try to be super smooth and consistent on all questions

Try to move constantly to mask any changes

Figure 4: Group 4 Manipulation (example)

Subjects

Subjects were recruited from undergraduate and graduate classes at a large public university in the southwest of the United States. Students are a good choice for this work as they are broadly similar to the population as a whole in terms of their technology familiarity and propensity to interact with online systems. In addition, they are good subjects since they are relatively homogenous, facilitating experimental control and internal validity (Dennis and Valacich 2001). We initially followed the standard practice of working with instructors to offer extra credit in their classes to subjects for participation. This serves to catch the attention of potential participants and motivate them to sign up. Then, we planned to use variable pay to incentivize the desired behavior (i.e., cheating on the test).

In our initial pilot tests, we had 15 questions on the test; five easy, five hard, and five impossible. Subjects were paid $5 to participate and then $1 per correct answer. The idea was to motivate the participants to get as many correct answers as possible to maximize their payment. After piloting the experiment with these incentives, we found that the rates of cheating were much lower than anticipated (5-10%). We reviewed the responses individually from the pilots and found that once participants encountered a hard or impossible question, they simply gave up. To get our cheating rates up to our target range (25-30%) needed for adequate sample sizes, we first upped the financial incentive. In the next pilot, instead of giving $1 per correct answer, we gave a $15 bonus for correctly answering all of the questions (100% correct). Our rates were still in the 5-10% range and the pattern of answers was similar – i.e., participants appeared to just give up.

We posit that the participants were being primarily motivated by the extra credit they were receiving just for participating and that this was overshadowing the financial incentives. Monetary incentives will induce effort if cognitive and motivation mechanisms are aligned (e.g., expected utility, goals, self-efficacy) (Bonner and Sprinkle 2002). The expected utility of cash is always greater than no cash, however, misaligned goals can overwhelm this effect (Bonner and Sprinkle 2002). In our case, the participant's goal was to gain extra credit in their course and thus was misaligned with the monetary incentive leading to lack of effort. To test this explanation, we ran another pilot where we specifically asked the instructors whose classes we recruited from to not offer extra credit for participation. Our sign-up rates were accordingly much lower, but the cheating rates went up to 30-35%[3]. Ultimately 282 subjects completed the study – 347 participated, but 65 used mobile devices to complete the survey so their data was not usable.

[3] This explanation is further supported by the difficulty we had in getting the initial batches of participants who received extra credit to come collect their cash payments. University policy required that we collect in-person signatures when providing payments and as such, the participants had to come to our office to receive payment. In the pilots with extra credit, multiple follow-up emails were required over the course of many days to get the students to come get their money. When they were only incentivized with monetary rewards, they appeared much more eager to come get their cash. Interestingly, however, many students under both reward mechanisms never did come get their money.

PRELIMINARY RESULTS

In analyzing the 282 subjects, we found some interesting trends in the preliminary results, however very little attained statistical significance. To investigate, a power analysis was performed to ascertain if the sample size was sufficient for the effects. Since the total number of subjects per cell is determined by the number needed to get a certain number of cheaters in that cell at the observed cheat rate, the power analysis was conducted on the subset of the sample that cheated. An estimate of the effect size for each metric on the pairwise t-tests for cheaters were calculated by taking the average of each pairwise mean difference and dividing by the pooled standard deviation for that metric (Cohen 1988). The mean overall effect size for cheaters was then calculated by taking the average across each of the measures level effect sizes (Cooper and Hedges 1993, p. 241). The overall estimated effect size value that was arrived at was 0.35. This is about halfway between the estimates of small (0.2) and medium (0.5) effect sizes for a t-test statistic (Cohen 1988, 1992), which was taken as evidence of face validity for the purposes of this power analysis. The power for the tests on the data that were collected was then calculated using the Cohen (1988) method and this estimate of effect size, yielding a power of 0.29. This indicates that the current experiment was substantially underpowered (Cohen 1988, 1992). A generally accepted value for sufficient power is 0.80 (Cohen 1992). To attain this level of power with our estimated effect size, we would need 102 cheaters per cell. Given our observed cheat rate of approximately 30-35%, we would thus need an overall sample size of approximately 300 per cell. A follow-up study has been conducted using Mechanical Turk in order to facilitate collecting so many subjects per cell. 2,500 subjects were collected: 7 cells at 300 per cell, plus a buffer to account for potential device filtering issues (e.g., a laptop with a touchscreen). We are currently analyzing this data.

CONCLUSION

This research in progress note reports on the design and execution of a study in HCI-based detection deception. The objective of the study was to examine the impact of knowledge of tracking and countermeasures on the neuro-motor changes detected when subject commit a malfeasant act. To examine this, an experimental context and design was required that would afford the subjects an opportunity to commit an un-sanctioned malicious act while tracking ground truth in an unobtrusive manner. While the initial data collection did not yield much in the way of significant results, the experimental context and design was found to be robust and potentially useful for others. Subsequent data collection has taken place and results will be reported in the future.

REFERENCES

1. ACFE. 2018. “2018 ACFE Report to the Nations,” Association of Certified Fraud Examiners. (http://www.acfe.com/rttn).
2. Bonner, S. E., and Sprinkle, G. B. 2002. “The Effects of Monetary Incentives on Effort and Task Performance: Theories, Evidence, and a Framework for Research,” *Accounting, Organizations and Society* (27:4), pp. 303–345.
3. Burgoon, J., Mayew, W. J., Giboney, J. S., Elkins, A. C., Moffitt, K., Dorn, B., Byrd, M., and Spitzley, L. 2016. “Which Spoken Language Markers Identify Deception in High-Stakes Settings? Evidence From Earnings Conference Calls,” *Journal of Language and Social Psychology* (35:2), pp. 123–157.
4. Coalition Against Insurance Fraud. 2015. “How Big Is $80 Billion? (Bigger than You Think!),” , December. (http://www.insurancefraud.org/80-billion.htm, accessed October 8, 2017).
5. Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, (2nd ed.), Hillsdale, N.J: L. Erlbaum Associates.
6. Cohen, J. 1992. “A Power Primer,” *Psychological Bulletin* (112:1), pp. 155–159.
7. Conti-Brown, P. 2017. “Wells Fargo Fake Accounts Scandal: Why It Might Not Survive | Fortune,” *Fortune*. (http://fortune.com/2017/08/31/wells-fargo-fake-accounts-scandal-2017-tim-sloan/).
8. Cooper, H., and Hedges, L. V. 1993. *The Handbook of Research Synthesis*, Russell Sage Foundation.
9. Dennis, A. R., and Valacich, J. S. 2001. “Conducting Experimental Research in Information Systems,” *Communications of the Association for Information Systems* (7:1), p. 5.
10. Eysenck, M. W., Derakshan, N., Santos, R., and Calvo, M. G. 2007. “Anxiety and Cognitive Performance: Attentional Control Theory.,” *Emotion* (7:2), p. 336.
11. Gorman, S. 2014. “Snowden Leaks Assailed in Senate Hearing on National Security,” *Wall Street Journal*. (https://www.wsj.com/articles/after-snowden-assessment-sees-leaks-as-among-top-us-threats-1391010295).
12. Hibbeln, M., Jenkins, J. L., Schneider, C., Valacich, J. S., and Weinmann, M. 2014. “Investigating the Effect of Insurance Fraud on Mouse Usage in Human-Computer Interactions,” in *35th International Conference on Information Systems, ICIS 2014*, Association for Information Systems.
13. Hibbeln, M., Jenkins, J. L., Schneider, C., Valacich, J. S., and Weinmann, M. 2017. “HOW IS YOUR USER FEELING? INFERRING EMOTION THROUGH HUMAN–COMPUTER INTERACTION DEVICES.,” *MIS Quarterly* (41:1).

14. Jenkins, J. L., Proudfoot, J. G., Valacich, J. S., Grimes, G. M., and Nunamaker Jr, J. F. 2019. “Sleight of Hand: Identifying Malicious Insider Threats through the Monitoring of Mouse-Cursor Movements,” *Journal for the Association for Information Systems* (Forthcoming).

15. Nunamaker Jr, J. F., Burgoon, J. K., and Giboney, J. S. 2016. “Special Issue: Information Systems for Deception Detection,” *Journal of Management Information Systems* (33:2), pp. 327–331.

16. Proudfoot, J. G., Jenkins, J. L., Burgoon, J. K., and Nunamaker Jr, J. F. 2016. “More Than Meets the Eye: How Oculometric Behaviors Evolve Over the Course of Automated Deception Detection Interactions,” *Journal of Management Information Systems* (33:2), pp. 332–360.

17. Su, L., and Levine, M. 2016. “Does ‘Lie to Me’ Lie to You? An Evaluation of Facial Clues to High-Stakes Deception,” *Computer Vision and Image Understanding* (147:Supplement C), Spontaneous Facial Behaviour Analysis, pp. 52–68.