Detecting Betrayers in Online Environments Using Active Indicators

Paola Rizzo[1](✉), Chaima Jemmali[1], Alice Leung[2],
Karen Haigh[2], and Magy Seif El-Nasr[1]

[1] Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA
{p.rizzo,m.seifel-nasr}@northeastern.edu, jemmali.c@husky.neu.edu
[2] Raytheon BBN Technologies, 10 Moulton Street, Cambridge, MA 02138, USA
{alice.leung,karen.haigh}@raytheon.com

Abstract. Research into betrayal ranges from case studies of real-world betrayers to controlled laboratory experiments. However, the capability of reliably detecting individuals who have previously betrayed through an analysis of their ongoing behavior (after the act of betrayal) has not been studied. To this aim, we propose a novel method composed of a game and several manipulations to stimulate and heighten emotions related to betrayal. We discuss the results of using this game and the manipulations as a mechanism to spot betrayers, with the goal of identifying important manipulations that can be used in future studies to detect betrayers in real-world contexts. In this paper, we discuss the methods and results of modeling the collected game data, which include behavioral logs, to identify betrayers. We used several analysis methods based both on psychology-based hypotheses as well as machine learning techniques. Results show that stimuli that target engagement, persistence, feedback to teammates, and team trust produce behaviors that can contribute to distinguishing betrayers from non-betrayers.

Keywords: Betrayal · Games as experimental methods · Deception
Espionage

1 Introduction

Betrayal of one's group is a phenomenon that occurs in many contexts and organizations [1–4], causing significant economic impacts. According to the Ponemon Institute [5], 874 insider incidents occurred in 2016 in the US, and of those 22% were criminal, costing USD 4.3 million. Therefore, it is important to develop techniques to identify persons who have engaged in such acts. While previous research investigated detecting betrayals within the act of betraying through anomaly detection or other methods (e.g., [3,6,7]), research detecting betrayers after the fact is sparse.

In this paper, we address this topic. In particular, we work with insider threat experts and psychologists who assume that persons who betray their

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 16–27, 2018.
https://doi.org/10.1007/978-3-319-93372-6_2

team or organization have emotional, logical thinking and habitual behaviors, similar to those discussed in [8,9], that are significantly different from those of people who do not betray their teams. We postulate that such behaviors can be evoked through stimuli in the environment producing a distinct fingerprint that is detectable by machine learning or statistical techniques.

In this paper, we focus on the emotional aspects, and expect that the individuals who have betrayed will feel guilty, anxious, trapped, and distant from the group [10]. Consequently, we expect betrayers to have less identification and trust with their teammates, and to exert less focus and diligence on their tasks, compared to other subjects. To investigate this hypothesis, we developed a novel methodological approach composed of multiple techniques. First, we implemented an online social game that allows participants to betray their group by sharing information with a competitor group. Second, in order to make our game a controllable environment with embedded stimuli that can cause subjects to behave in certain ways that can be detectable of malicious intent, we used a technique similar to Sasaki's work [7], according to whom psychological triggers that heighten anxiety in malicious insiders cause them to carry out specific behaviors of deleting evidence and stopping further malicious activities. For instance, a stimulus suggesting that file-searching behaviors may be under surveillance is likely to be ignored by a normal subject engaged in work-related searches, but may cause a malicious subject engaged in espionage to cease certain activities [11]. We developed 13 psychological stimuli called "Active Indicators" (AIs), designed to evoke behaviors that can distinguish betrayers in an online environment. A few of these stimuli, rather than being obtrusive, are integral parts of the background activities, such as stimuli embedded in team text chat, opportunities to react to prompts and cues during the game.

Our work provides the following contributions. First, it presents a novel methodological approach to investigate a set of manipulations to detect betrayers behaviorally after the act of betrayal. The method is based on a game that embeds AIs developed according to previous research. Lessons about the design of the game as well as the utility of previous research can provide a good step for researchers interested in studying this topic within other contexts. Second, we discuss the AIs and the resulting patterns of behaviors and their power in detecting betrayers. We found AIs that target engagement, persistence, feedback to teammates and team trust to be among the most significant and further work is needed to validate these results in real-world environments.

2 Related Work

Computer-based insider threats have been a subject of study for years. Some works focus on anomaly detection, i.e. automated ways using machine learning to distinguish suspicious activities from regular ones [12,13], while other works investigate the utility of eliciting spying behaviors by means of "honeypots", custom-built information system resources (e.g., special files) that can attract and reveal potential insiders [14]. However, none of these works detect insider threats after the fact.

Some works (e.g., [3,6]) use psycho-social and behavioral indicators as antecedents of insider threat activities to assess the chances that an individual will perform specific behaviors. Sasaki [7] assumes that malicious insiders are anxious about their identity being revealed and thus psychological triggers should heighten their anxiety and cause-specific behaviors that will reveal them. We share his hypothesis, however, we focused on emotional stimuli beyond anxiety, including guilt, distance from the group, feeling trapped. We also looked at behavioral patterns other than those concerning insider threats, e.g., willingness to carry out extra work and identification with the team.

Other researchers focus on deceptive communication, looking at nonverbal behaviors such as facial expressions [15], or at linguistic patterns in chats. For instance, Niculae et al. ([16]) studied dyadic communication in an online game where players break alliances through betrayal, but they focused on linguistic cues that foretell betrayal rather than communication patterns of betrayers. Ho and Warkentin [17] developed a game to examine the trustworthiness of spies as measured by teammates engaging in computer-mediated communication. Additionally, Ho et al. [18] used Support Vector Machines to classify deceivers and non-deceivers based on cues in chat data, and found that cues related to time-lag, social attitude and negation in text can discriminate between deceivers and non-deceivers. While such research is relevant, our work uses behavioral measures rather than relying on human-perceived trustworthiness or the contents of chat data.

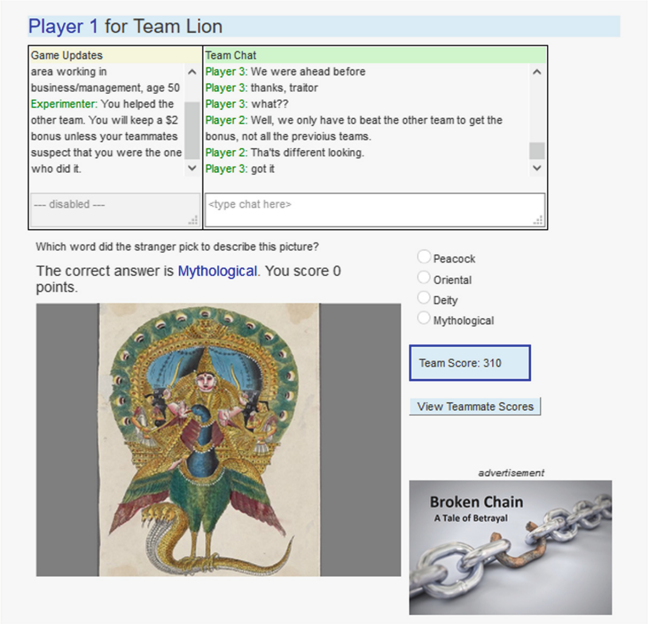
Several works show that deceivers may appear more submissive than truth-tellers when their primary goal is to evade detection (e.g., [19]). However, other research shows that this pattern is reversed when deceivers need to persuade others of their credibility, and thus tend to argue aggressively while simultaneously trying to avoid being detected, a behavior called persuasive deception [20]. In such a case, deceivers may display more dominance, using verbal and non-verbal communication that makes them appear confident [20]. Also, the style of deception can change according to whether the recipient is acquiescent or suspicious [21]. As discussed later in the paper, the textual communicative behavior of betrayers in our game is somewhat similar to that of persuasive deceivers.

3 The Game: ESP

We designed a simple guessing game (see Fig. 1), inspired by Von Ahn and Dabbish [22], lasting about 50 min, where a team plays against another. The goal of the game is to guess the gender, age, location and occupation of a stranger, based on accumulated information about the stranger's reaction to a series of pictures. Each team selects a stranger to be guessed by the opponent team, and then earns points when its members correctly answer questions about its own target stranger. Teammates collaborate to find the right answer by communicating through chat. A game session lasts five rounds, each including 3 pictures of art and 2 questions per picture (first "Which word did the stranger pick to describe this picture?", and then "Did they like the picture?" or "What was their favorite thing about the picture?"). After each round, the scores of each team are revealed. After the last round, 4 high-point value questions about the stranger's demographics, and the final scores of each team are revealed. We let participants play with pre-scripted bot team members, against an imaginary opponent team, to maintain control and comparability across teams (the automated nature of team members and the opponent team was not disclosed to the subjects).

4 Experimental Manipulations and AIs

We had two experimental conditions: the control group played the game with no opportunity to betray their team, while the experimental group was given a message (shown on screen at the end of the first round of the game) asking them if they would secretly pass information to the opponent team about the target stranger and receive a $2 bonus payment in return. It was then up to them to betray their team or not by answering "yes" or "no". After the choice was made, the game announced: "One member of your team was offered money to tell the



area working in business/management, age 50
Experimenter: You helped the other team. You will keep a $2 bonus unless your teammates suspect that you were the one who did it.

Player 3: We were ahead before
Player 3: thanks, traitor
Player 3: what??
Player 2: Well, we only have to beat the other team to get the bonus, not all the previoius teams.
Player 2: Tha'ts different looking.
Player 3: got it

Which word did the stranger pick to describe this picture?

The correct answer is Mythological. You score 0 points.

○ Peacock
○ Oriental
○ Deity
○ Mythological

Team Score: 310

Fig. 1. A screenshot of the ESP game, comprising a chat window (top), a large window where the picture is displayed (bottom left), the question about the picture (below the chat window), the team's score (below the answer options), and a window containing a variable picture (bottom right) used as a small advertisement area for priming and psychological stimuli.

other team the gender of the stranger you picked." In case of betrayal, the text continued with "That player accepted the offer, but they will lose the money if the rest of the team suspects them. Later, the team will vote on who sold the information", otherwise "That player declined the offer and stayed loyal to your team". Hence, we had 3 groups of subjects: controls, betrayers, and decliners.

The game was designed to evoke anxiety and guilt, by showing (a) the negative consequences of discovery, where participants were told they would lose the bonus if their teammates suspected they were the betrayer, and (b) the negative impact of betrayal on the player's teammates, by letting the latter react strongly to the announcement that someone had betrayed the team and produce negative comments about the betrayal during the rest of the game.

At the end of the experiment, we sent participants a debriefing statement that included: "We used this experiment to understand what happens when people put their individual gain before their team's. You were all given the maximum bonus as a reward for your participation". The whole experimental protocol was approved by the IRB.

Based on betrayers' expected behaviors, we developed several AIs (Active Indicators), and corresponding measurements, grouped according to the affective aspect targeted. Our goal is to examine if such AIs when embedded in the environment produce behaviors that can easily discriminate betrayers from the two other groups (controls and decliners). Many AIs were mapped onto several stimuli or "prompts" to provide repeated measures of the same AI while avoiding too much obvious repetition during the experiment. Furthermore, some of the stimuli were embedded in the game rather than being introduced as additional events (for instance, this was the case for AI-Timer and AI-Disengage listed below).

The targeted affective states and corresponding AI stimuli and expected behaviors are as follows:

- **Reluctance to be identified with the group, as a result of emotional distance**
  - *AI-Criticism-of-Group*: Exposure to criticism of the group - *Example stimulus*: presentation of negative information about the team - *Expected behavior*: agreeing with criticism or declining to defend the team from criticism
  - *AI-Assessment-of-Group*: Ask for group assessment - *Example stimulus*: game poll asking the participant to "Please share your impressions about your team" - *Expected behavior*: negative and neutral assessment of the team
- **Less regard and trust for those betrayed, as a result of projection of own lack of trustworthiness and emotional distance**
  - *AI-Peer-Feedback*: Present opportunity to get/give peer feedback - *Example stimulus*: game message asking "Would you like to get feedback from a teammate about your participation? (Y/N)" - *Expected behavior*: "no"
  - *AI-Check-Peer-Performance*: Present opportunity to check peer performance - *Example stimulus*: a button on that can pop up information

about each team member's contribution to the team score - *Expected behavior*: click on the button
  - *AI-Peer-Trustworthiness*: Ask for an assessment of peer trustworthiness - *Example stimulus*: game message asking "How trustworthy are your teammates?" - *Expected behavior*: low assessment of trustworthiness of teammates
  - *AI-Follow-group*: Provide opportunity to follow group advice - *Stimuli*: presentation of teammate advice throughout the entire task - *Expected behavior*: choose answers that disagree with teammate advice
- **Lack of persistence, as a result of being trapped**
  - *AI-Slack-Off*: Present opportunity to slack off on work - *Stimuli*: opportunities to contribute to team success through chat - *Expected behavior*: low participation in team discussion
- **Impulsive risk taking as a result of anxiety about discovery**
  - *AI-Wager*: Present opportunity to wager - *Stimulus*: an opportunity to wager on an outcome only marginally under the participant's control - *Expected behavior*: decision to wager
- **Does not devote full attention to job, as a result of anxiety about discovery**
  - *AI-Timer*: Present a time-limited window to complete routine work - *Stimuli*: presentation of deadlines for submitting work, as opportunities to submit answers before a timer runs out - *Expected behavior*: not submitting an answer before a timer runs out
  - *AI-Extra-Info*: Present opportunity to gather information to improve probability of success on a task - *Example Stimulus*: presentation of an opportunity to read an extra essay about the stranger - *Expected behavior*: refusal to read the extra essay
  - *AI-Shorter-Game*: Present choice between a smaller task for less reward or a larger task for more reward - *Example Stimulus*: presentation of the poll "Would you prefer to play a longer game (with the opportunity to earn more points) or a shorter game?" - *Expected behavior*: opt for a shorter game
  - *AI-Disengage*: Present opportunities to disengage from task - *Stimuli*: presentation of non-obtrusive task activities throughout the entire game (team text chat, opportunities to react to prompts and cues during the game, and other game content) - *Expected behaviors*: longer time taken to react to game prompts, higher number of times a participant neglects to respond to a game prompt, higher frequency and duration of non-game browser window activity
  - *AI-Cognitive-Challenge*: Present a cognitive challenge - *Stimulus*: a quick test of short memory recall - *Expected behavior*: failure to correctly recall any sequence, caused by not engaging with the test.

4.1 Subjects

We recruited a total of 348 subjects from Amazon Mechanical Turk, with the requirement that participants need to be US residents: 52% males, 48% females, and 88% who attended college; the age mode was 25–29 years, with a frequency of 25%. The compensation was of \$5 + \$2 bonus, set on the basis of several tests aimed at finding a good balance between the percentage of subjects who accepted to betray their team and the percentage of those who did not. We excluded 115 subjects from the analyses because they did not answer the post-game survey, or because they expressed a belief that their teammates were bots or experimenters. They could express such belief either during chat, or as part of their free-text responses to questions about the team during the game, or in the post-game survey. In fact, we assumed that participants would not develop the same social and emotional reactions to betrayal of presumed computer controlled entities or experimenters as they would for presumed human teammates. 76 of our participants were betrayers, 74 were decliners and 83 were controls.

4.2 Data Collected

We collected 2 types of data: behavioral logs of actions in the game and self-report measures. The behavioral data were in the form of time-stamped entries for what the participant saw and did: game content, text chats, button clicks, participant score, etc. As for the self-report measures, immediately after the game each participant was asked to complete a short demographic survey, as well as the validated survey PANAS (Positive Affect Negative Affect) [23], that gauges the respondent's affective state.

We preprocessed the behavioral data to develop measures ready for analysis. There were some variations in the AIs used, in that some AIs had both multiple signals and multiple time segment detection points, and some AIs were continuous measures, others were discrete, and some were human coded rather than automatically labeled. For continuous measures, we aggregated them at three different points: segment 0 (before the opportunity for betrayal), segment 1 (after the betrayal decision point but before the priming cue), and segment 2 (after both the betrayal decision point and the priming cue). For control participants who did not have a betrayal decision point, the corresponding time point in the game task was used. We then normalized the measures across the different segments with the baseline established as the behavior for segment 0 (the AI before inducement). For discrete measures, concatenation was used to make sure to record before and after the stimuli.

4.3 Analysis Methods and Results

Regarding the self-report measures (see Table 1 below), PANAS "Guilt" produced a stronger effect on betrayers, with a significant difference between them and both controls and decliners (one-tail t test $p < 0.0001$). As for anxiety, the "Afraid" and "Scared" measures of the PANAS scale showed some significant differences between betrayers and other subjects (one-tail t-tests $p < 0.05$). We also found a significant difference between betrayers and other subjects regarding the PANAS "Ashamed" measure (one-tail t-tests $p < 0.01$). This is a likely effect of the negative reactions of the team members to the betrayal.

Table 1. Statistics about self-reported measures.

Variable	Group	Statistics	Variable	Group	Statistics
"Guilt"	Betrayers	$\mu = 3$, $\sigma = 1.21$	"Ashamed"	Betrayers	$\mu = 2.82$, $\sigma = 1.14$
"Guilt"	Controls	$\mu = 2.13$, $\sigma = 0.73$	"Ashamed"	Controls	$\mu = 2.25$, $\sigma = 0.60$
"Guilt"	Decliners	$\mu = 2.32$, $\sigma = 0.78$	"Ashamed"	Decliners	$\mu = 2.35$, $\sigma = 0.83$
"Afraid"	Betrayers	$\mu = 2.46$, $\sigma = 0.99$	"Scared"	Betrayers	$\mu = 2.46$, $\sigma = 0.97$
"Afraid"	Controls	$\mu = 2.20$, $\sigma = 0.62$	"Scared"	Controls	$\mu = 2.20$, $\sigma = 0.62$
"Afraid"	Decliners	$\mu = 2.23$, $\sigma = 0.63$	"Scared"	Decliners	$\mu = 2.21$, $\sigma = 0.69$

We used two approaches to analyze the effect of AIs: (1) theory-based detector rules computed on the measures of single AIs and (2) theory-agnostic Machine Learning detector rules computed on the measures of single and multiple AIs. Due to space limitations, we will only discuss the results of AIs that were statistically significant or discriminative.

For (1), we defined and tested simple detector rules based on theory expectations about how betrayers' behavior should differ from controls' on each AI sub-measure. For example, for sub-measures that simply detect whether a participant responded to a prompt, a simple rule was "Betrayers do not respond, other participants do respond." For sub-measures that were scales (e.g. degree of positivity in response) or continuous (e.g. amount of time spent with the game window not activated), a "cut-off" value was selected based on criteria to balance between differentiation (TP/FP ratio) and detection (TP). We used the same process to set the cut-off value for each sub-measure. For this process, an initial cut-off value was selected to divide participants into betrayers vs controls or decliners by calculating the cut-off value such that 20% of betrayers were included. Then, the cut-off value was adjusted in the direction of increasing TP until a local maximum of TP/FP was found. We tested the performance of these rules for both the betrayer/control separation and the betrayer/decliner separation. The best discriminations between betrayers and controls produced by this method are shown in Table 2.

For (2), we used Machine Learning (ML) classification methods with the three experimental conditions (betrayers, decliners, and controls) as labels and AI measures as features. This approach estimates how much discriminative power an AI provides, agnostic to whether the rule follows psychological theory, and can screen both single and composite indicators (the latter made up of two or more individual AIs) to test whether they would provide more discrimination in combination. We also included demographics and post-game surveys as features to see if they could have discriminative power. We ran eight types of algorithms provided by the Weka ML library [24]: Functions (Support Vector Machines using Pearson VII Universal Kernel), Lazy Models (IBk (kNN classifier), KStar (instance-based learner)), Rules (JRip (RIPPER), Ridor (RIpple-DOwn Rule Learner)), Trees (FT (Functional Trees), J48 (C4.5 decision tree)), and Misc (VFI (Voting feature intervals)). The input feature vector is composed of 15 features (13 AIs plus the data from the PANAS and another survey).

This analysis enabled us to estimate whether our behavioral AI measures are more or less discriminative than individual characteristics or self-reported feelings. Each feature is composed of multiple sub-measures. We developed models that used one feature and models that used 381 combinations of features. We examined the results for both the f measure (as a function of precision and recall) and the TP/FP ratio. For each model type, we computed a random 75%/25% test/train split, trained the model and then computed the f measure and TP/FP ratio on the resulting confusion matrix. Table 2 lists the results of the best performing single-feature and multi-feature classifiers for betrayer/control separations.

Table 2. Summative table of best performing *AIs* using Theory-based rules and Machine Learning classifiers

Classifier type	AI	Results
Theory-based rules	*AI-Disengage*	TP/FP ratio = 6.09
	AI-Slack-Off	TP/FP ratio = 5.19
ML single-feature classifiers	*AI-Peer-Feedback*	TP/FP ratio = 6.8, f measure = 0.50
	AI-Peer-Trustworthiness	TP/FP ratio = 2.72, f measure = 0.51
	AI-Slack-Off	TP/FP ratio = 2.66, f measure = 0.37
	AI-Disengage	TP/FP ratio = 3.74, f measure = 0.28
ML multi-feature classifiers	all the best ones included: *AI-Peer-Trustworthiness*	TP/FP ratios > 2.4, f measures > 0.50
	15-feature model	TP/FP ratio = 2.93, f measure = 0.54
	the best 15-feature model included: *AI-Check-Peer-Performance* *AI-Peer-Trustworthiness* *AI-Shorter-Game*	TP/FP ratio = 4.16, f measure = 0.66

The most discriminative sub-measure was *AI-Disengage-B*, the number of text chats typed by the participant: it achieved a TP/FP ratio of 6.24 with TP = 30%. In fact, after the inducement to betray, the number of chats was significantly higher for betrayers than for decliners and controls (segment 1 - after the betrayal decision point but before the priming cue- betrayers: $\mu = 23.49$, $\sigma = 13.99$; decliners: $\mu = 16.58$, $\sigma = 11.33$; controls: $\mu = 12.83$, $\sigma = 9.39$; two-tails t-tests $p < 0.01$).

The best performing classifier varied widely across the different behavior measures. All eight models yielded the best result for some of the input vectors. JRip, FT and J48 tended to do better. See Table 3 for the details.

Even though the betrayers of our experiments were not requested to actively engage in sustained deceptive communication, they may have used communication strategies analogous to those of deceivers in prior studies, in that they chatted more than other groups ([20]), and may have produced effects similar

Table 3. Performance of ML algorithms

ML type	ML technique	Count of testdata best model (TP/FP ratio Yes)	Count of testdata best model (FMeas)
Function	SVM	29	30
Lazy	IBk	16	24
Lazy	KStar	30	27
Misc	VFI	46	25
Rules	JRip	90	122
Rules	Ridor	21	12
Trees	FT	76	67
Trees	J48	73	74
Grand Total		381	381

to those found by Anolli et al. [21] when "lying to a suspicious recipient". In fact, the strong negative reactions of the teammates to the announcement of the betrayal, and the anxiety caused by the risk of being caught, may have caused betrayers to actively attempt to persuade teammates about their innocence by showing an active participation to the game so as to be seen as good team members.

We also informally examined the chats to identify the possible pragmatic strategies used by betrayers vs decliners and controls, but could not find qualitative differences between the three groups of subjects, in that all groups seem to pursue the same communicative goals but with different frequencies. We could not tell the intentions behind the betrayers' increased frequency of such communicative goals without another study that includes interviews or betrayers' reflection on their behaviors.

5 Conclusions and Future Work

In this work, we aimed to study betrayers by placing "Active Indicators" (AIs) in the environment to elicit indicative responses. Our work is the first to use a game to explore emotional indicators as a way to detect betrayers using online behaviors after the act of betrayal. The game provided us with a controllable environment, including replicable teammates, where we could continuously monitor the subjects' behaviors to deduce the effects of stimuli on them. We measured the detector signals of each AI by collecting behavioral data in the form of time-stamped entries for what the participants saw and did. Our results show that some AIs (those that target engagement, persistence, feedback to teammates and team trust) have a promising discriminatory power, both taken singularly and combined with other AIs. One AI related to engagement was the number of chats, which was much higher for betrayers than controls and decliners, confirming other results in the literature [20,21]. In our case, betrayers did "role-playing": they probably produced more chats than decliners and controls to pretend that they were good team members, by participating more in the discussions. For future work, we aim to conduct more qualitative analyses to understand the communicative strategies that betrayers used as opposed to others. We also plan to use other games as experimental environments to analyze the effects of AIs on betrayal behaviors controlled by habits and logical reasoning. These results are promising and follow on studies should take this further to investigate the effect of the successful AIs in different contexts.

Acknowledgments. The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract number 2016-16031100003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

1. Herbig, K.L., Wiskoff, M.F.: Espionage against the united states by American citizens 1947–2001. Technical report, Defense Personnel Security Research Center Monterey CA (2002)
2. Cummings, A., Lewellen, T., McIntire, D., Moore, A.P., Trzeciak, R.: Insider threat study: illicit cyber activity involving fraud in the us financial services sector. Technical report, Carnegie Mellon University, Software Engineering Institute, Pittsburgh, PA (2012)
3. Kont, M., Pihelgas, M., Wojtkowiak, J., Trinberg, L., Osula, A.M.: Insider threat detection study. NATO CCD COE, Tallinn (2015)
4. Carter, M.: Massively multiplayer dark play: Treacherous play in eve online. The Dark Side of Game Play. Routledge, London (2015)
5. Ponemon Institute: 2016 cost of insider threats. Technical report, Traverse City, MI, USA (2016)
6. Greitzer, F.L., Paulson, P., Kangas, L., Franklin, L.R., Edgar, T.W., Frincke, D.A.: Predictive modelling for insider threat mitigation. Pacific Northwest National Laboratory, Richland, WA, Technical report, PNNL-65204 (2009)
7. Sasaki, T.: A framework for detecting insider threats using psychological triggers. JoWUA **3**(1/2), 99–119 (2012)
8. Lane, J.D., Wegner, D.M.: The cognitive consequences of secrecy. J. Personal. Soc. Psychol. **69**(2), 237 (1995)
9. Kelly, A.E.: The Psychology of Secrets. Springer, Heidelberg (2002). https://doi.org/10.1007/978-1-4615-0683-6
10. Charney, D.: True psychology of the insider spy. Intell.: J. US Intell. Stud. **18**(1), 47–54 (2010)
11. Intelligence Advanced Research Program Agency (IARPA): Iarpa scite program (2015)
12. Rashid, T., Agrafiotis, I., Nurse, J.R.: A new take on detecting insider threats: exploring the use of hidden markov models. In: Proceedings of the 2016 International Workshop on Managing Insider Security Threats, pp. 47–56. ACM (2016)

13. Bindu, P., Thilagam, P.S.: Mining social networks for anomalies: methods and challenges. J. Netw. Comput. Appl. **68**, 213–229 (2016)
14. Spitzner, L.: Honeypots: catching the insider threat. In: Computer Security Applications Conference, 2003. Proceedings. 19th Annual, pp. 170–179. IEEE (2003)
15. Peled, N., Bitan, M., Keshet, J., Kraus, S.: Predicting human strategic decisions using facial expressions. In: IJCAI, pp. 2035–2041 (2013)
16. Niculae, V., Kumar, S., Boyd-Graber, J., Danescu-Niculescu-Mizil, C.: Linguistic harbingers of betrayal: a case study on an online strategy game. arXiv preprint arXiv:1506.04744 (2015)
17. Ho, S.M., Warkentin, M.: Leader's dilemma game: an experimental design for cyber insider threat research. Inf. Syst. Front. **19**(2), 377–396 (2017)
18. Ho, S.M., Liu, X., Booth, C., Hariharan, A.: Saint or sinner? language-action cues for modeling deception using support vector machines. In: Xu, K., Reitter, D., Lee, D., Osgood, N. (eds.) Social, Cultural, and Behavioral Modeling. LNCS, vol. 9708, pp. 325–334. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-319-39931-7_31
19. Burgoon, J.K., Dunbar, N.E.: An interactionist perspective on dominance-submission: interpersonal dominance as a dynamic, situationally contingent social skill. Commun. Monogr. **67**(1), 96–121 (2000)
20. Dunbar, N.E., Jensen, M.L., Bessarabova, E., Burgoon, J.K., Bernard, D.R., Harrison, K.J., Kelley, K.M., Adame, B.J., Eckstein, J.M.: Empowered by persuasive deception: the effects of power and deception on dominance, credibility, and decision making. Commun. Res. **41**(6), 852–876 (2014)
21. Anolli, L., Balconi, M., Ciceri, R.: Linguistic styles in deceptive communication: dubitative ambiguity and elliptic eluding in packaged lies. Soc. Behav. Personal.: Int. J. **31**(7), 687–710 (2003)
22. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 319–326. ACM (2004)
23. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: the panas scales. J. Personal. Soc. Psychol. **54**(6), 1063 (1988)
24. Frank, E., Hall, M.A., Witten, I.H.: Weka 3: Data mining software in java (2017). http://www.cs.waikato.ac.nz/ml/weka/