If you betray your teammates, do you think you can be spotted? *

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

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

Abstract. Detecting betrayers and liars in virtual environments is a topic of interest to many organizations and has spurred research for years. Here we address this problem by testing how well emotional Active Indicators (unobtrusive, deliberately introduced stimuli) trigger behaviors that distinguish betrayers. We focus on a theoretical framework about mental states that result from betrayal and that may affect subsequent behavior. To this aim, we developed an online chat-based game where participants are given a choice to betray their team by providing information to an opponent team. We embedded many automatically deployed active indicators in the game. Then we used statistical and machine learning techniques to develop models to discriminate between betrayers (people who chose to betray), non-betrayers (people who chose not to betray), and controls (people who were not given a choice to betray) based on the behavioral responses to stimuli. We also looked at the influence of demographics, personality and other factors on players' choice to betray and their behaviors. Results show that betrayers engaged in chatting more than other groups, which suggests that they may use deceptive communication strategies analogous to those described in previous work. In addition to discussing results, in this paper we are also presenting the use of games as a method to investigate and deeply examine deceptive behavior in a controllable manner.

Keywords: Behavior Modeling, Deceptive Communication, Online games.

* 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.

1 Introduction

Online insider threats under the form of information leakage can produce significant harm to public and private organizations, and are difficult to identify and distinguish – a problem that spurred much research over the past few years (e.g. [Charney 2010], [Ho & Warkentin 2017], [Sasaki 2012], [Spitzner 2003]). A possible way to tackle this problem is based on "Active Indicators" (AIs), stimuli designed to evoke responses that are more characteristic of malicious subjects than normal subjects: e.g., a stimulus that suggests that certain file-searching behaviors may be noticed 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 [IARPA 2015].

Our goal was to design emotion-related AIs according to the likely psychological state of betrayers (that may feel guilt and anxiety), and apply the AIs to a game, where we can reproduce some features of a group setting that is akin to real-life and where we can invite a certain number of participants to betray their team by sharing information with another competing team. In this paper we show that the emotional impacts of betrayal could develop relatively quickly against a simple background task offered by a 1-hour long game. Additionally, the paper discusses the game we created, shows the results of building a behavioral model (with respect to non-betrayers, betrayers engage in more chat, are more likely to rate their teammates as less trustworthy, and feel more guilty and anxious), and describes some similarities between betrayers' and deceivers' communicative strategies.

2 Related Work

As mentioned above, a possible approach for detecting spies is by deliberately placing designed stimuli in the environment that can induce indicative responses from persons engaged in insider threats or espionage [Sasaki 2012]. The idea behind such an approach is that the practice of espionage/deception/betrayal/spying leaves recognizable mental effects on the actor's emotions, habits, and logical reasoning, and that these effects can be revealed by the actor's response to certain stimuli targeted at these emotional, habitual and logical behaviors. Our work focuses on Active Indicators that cause emotional behaviors: taking inspiration from [Charney 2010], we have designed stimuli that are expected to trigger guilt and anxiety in betrayers and cause them to perform behaviors that non-betrayers are less likely to perform.

We chose to embed the AIs in an online game, because game environments can be well suited to study a wide range of human behaviors, including attitudes and normative behaviors [Ajzen and Fishbein 1970], visual attention [Seif El-Nasr and Yan 2006], [Badler and Canossa 2015], emotions and motivation [MacDowell and Mandler, 1989], [Adbuhamdeh et al. 2015], personality [Canossa et al., 2015], social psychology [Blascovich et al. 2002], and trust [Seif El-Nasr et al. 2014].

As we will see later on, the communicative behavior of betrayers in our game in some respects look similar to that of deceivers. There is a substantial amount of work about deceptive communication that shows, among other things, that deceivers may appear more submissive than truth-tellers when their primary goal is to evade detection (e.g., [Burgoon & Dunbar, 2000]). Research also shows that this pattern is reversed when deceivers need to persuade others of their credibility, and they tend to argue in favor of the position they are supporting while simultaneously trying to avoid being detected, a type of lying named persuasive deception. In such a case, deceivers may display more dominance, using verbal and nonverbal communication that let them look confident (e.g. [Dunbar et al. 2014]). Also, the style of deception can change according to whether the recipient is acquiescent or suspicious [Anolli et al. 2003].

There are also specific text-based cues about deception that have been discussed in previous literature. For instance, Ho et al. [2016] used Support Vector Machines to classify deceivers and non-deceivers based on such cues in chat data, where they discussed how cues related to time-lag, social attitude and negation in text can discriminate between deceiver and non-deceivers.

In our work, we do not analyze the actual chat, but rather look at the behavioral patterns over time. Thus, we have been able to identify clearly and empirically the patterns described by Dunbar et al. [2014]. In the future we aim to further analyze the chat to see if we find similar patterns to the reported cues in Ho et al.'s work.

3 Hypotheses and experiment

At a high level, we hypothesize that betrayers (people who were asked and agreed to share information with an opposing team, thus betraying their own team) would exhibit less identification and trust with their teammates, and less focus and diligence on their task, compared to decliners (people who declined to share information with an opposing team) and controls (people who were not given a choice to betray). Based on previous work [Charney 2010] and Subject Matter Experts involved in our team, we developed a set of target behaviors for 18 AIs (or stimuli embedded within the game), and we hypothesize that the behaviors triggered by these AIs would discriminate between the three groups (betrayers, decliners and controls). A couple of example AIs with the same target behavior are the following:

- **AI Stimulus**: AI-24 – Present opportunity to gather information to improve probability of success on a task; **Detector Signal**: How much information is sought (betrayers should be less likely to seek the information); **Target Behavior**: Doesn't devote full attention to job, as a result of anxiety about discovery
- **AI Stimulus**: AI-11 – Expose to criticism of group; **Detector Signal**: How much objection to criticism is exhibited (betrayers should exhibit less objection to criticism); **Target Behavior**: Reluctance to be a representative of, or be identified with the group, as a result of emotional distance in order to act against group interests

To test our hypotheses we designed a simple guessing game, lasting about 50 minutes, presented as a team against team contest, with members of the winning team earning a bonus payment. The goal of the game is to guess the gender, age and occupation of a stranger. Each team decides the stranger to be guessed by the opponent team, and then earns points when its members correctly answer questions about its own target stranger. Teammates communicate through text chat to share their theories and help each other answer questions correctly. The team consists of 1 human player and 3 completely prescripted bots, and the opponent team is fake to maintain experimental control and comparability across teams (the fake nature of the team members and of the opponent team is not disclosed to the human player).

A game session lasts five rounds, each including 3 pictures of art and 2 questions per picture ("Which word did the stranger pick to describe this picture?" and "Did they like the picture?" or "What was their favorite thing about the picture?"). After each round, the team score of the opponent team is revealed, and after the last round there are 4 high-point value questions about the stranger's demographic characteristics, and the final team score of the opponent team is revealed.

Control group participants play the game with no opportunity to betray their team, while inducement group participants are offered, at the end of the first round of the game, a chance to receive a $2 bonus payment in return for secretly passing information to the opponent team about the latter's target stranger.

We emphasized the negative consequences of discovery, by telling participants that they could only keep the payment bonus if the other members of their team did not identify them as the betrayer, and the negative impact of betrayal on one's teammates participants, by letting the latter react more strongly to the announcement that someone had betrayed the team.

We logged time stamped entries for what the participant saw (virtual screen) and did: game content, text chats, button clicks, participant score, loss of window focus (e.g., participant was doing something else on their computer during game play). We ran the game on the Volunteer Science platform (https://volunteerscience.com/), and published it as work-for-hire on the Amazon Mechanical Turk crowdsourcing platform (subjects received a $5 reward plus a $2 bonus).

Participants were asked to complete a short survey immediately following the game to get demographic data and gauge their reactions to the game and emotional state. After a few days, participants were asked to answer the IPIP-NEO (Big Five personality traits and facets) and the NPI (Narcissist Personality Inventory) surveys for an extra $2.

4 Results and discussion

We recruited 348 participants, of which 76 betrayers, 83 controls and 74 decliners. 115 participants were removed from the analysis pool because they did not answer the post-game survey, or because, during chat or as part of their free-text responses to questions about the team during the game or in the post-game survey, they expressed a belief that their teammates were bots or experimenters. In fact we assume 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.

We computed correlations between behavior measures of AIs effects, ran t-tests, and developed and tested machine learning (ML) detector rules. The latter provide an estimate of how much discriminative power an active indicator's behavior signal provides, agnostic to whether the rule follows psychological theory, and can screen composite indicators, made up of two or three individual AIs, to test whether they would provide more discrimination in combination. To develop ML rules, we ran several types of algorithms provided by the Weka ML library (http://www.cs.waikato.ac.nz/~ml/weka/): SVM using Pearson VII Universal Kernel, Bayesian models, lazy models (IBk and KStar), rules (JRip, Ridor), Functional Trees, C4.5 decision tree, and Voting Feature Intervals. We tested the performance of ML generated detector rules for both the Betrayer/Control and Betrayer/Decliner separation, and we included demographic, personality, and post-game surveys as classifier inputs. This enabled us to estimate whether our active indicator behavior measures are discriminative compared to individual characteristics or self-reported feelings.

The most discriminative active indicator was AI7 – Team trustworthiness (Voting Feature Intervals F measure for Betrayers vs Controls: 0.645), i.e. the opportunity to rate team and teammate trustworthiness: (a) *betrayers were more likely to rate their team and teammates low on trustworthiness*. This behavior may be due to the psychological phenomenon of “projection” (e.g., believing that others are not trustworthy because oneself is not trustworthy). The next most discriminative active indicator was AI27 – Engagement (Support Vector Machines F measure for Betrayers vs Decliners: 0.445), that included the teammate chat, both spontaneous and in response to the opportunity to publicly communicate to the team. Specifically: (b) *betrayers were more likely to engage in a high volume of chat*.

As for the statistical tests, the PANAS “Guilt” measure was slightly higher in betrayers as expected (betrayers: M = 3, SD = 1.21; controls: M = 2.13, SD = 0.73; decliners: M = 2.32, SD = 0.78), with a significant difference between them and both controls and decliners (one-tail t test $p < 0.0001$). Also the “Afraid” and “Scared” measures of the PANAS scale showed significant differences between betrayers and other subjects as expected (e.g., for “Afraid”, we have betrayers: M = 2.46, SD = 0.99; controls: M = 2.20, SD = 0.62; decliners: M = 2.23, SD = 0.63; one-tail t test $p < 0.05$).

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 much more than the other groups, and seemingly exhibited a more emotionally strong chat and team-oriented attitude. In fact, the strong negative reactions of the teammates to the announcement of the betrayal, and the risk of being caught, may have caused betrayers to actively attempt to persuade teammates about their innocence (“persuasive deception”, see [Dunbar et al. 2014], and produced effects similar to those found by [Anolli et al. 2003] when “lying to a suspicious recipient”.

This paper makes two concrete contributions: 1) from an empirical point of view, it confirms previous work, showing that betrayers engage in more chat and are more likely to rate their teammates as less trustworthy; 2) from a methodological point of view, it shows the use of games as a method to deeply analyze betrayal and deception like behaviors. For future work, we plan to do more analysis on chat data influenced by such works and by the work by [Ho et al. 2016].

References

1. Abuhamdeh, S., Csikszentmihalyi, M., Jalal, B.: Enjoying the possibility of defeat: Outcome uncertainty, suspense, and intrinsic motivation. Motivation and Emotion, 39 (1), 1-10 (2015).
2. Ajzen, I., & Fishbein, M.: The prediction of behavior from attitudinal and normative variables. Journal of Experimental Social Psychology, 6 (4), 466-487 (1970).
3. Anolli, L., Balconi, M., Ciceri, R.: Linguistic styles in deceptive communication: Dubitative ambiguity and elliptic eluding in packaged lies. Social Behavior and Personality, 31, 687-710 (2003).
4. Badler, J., Canossa, A.: Anticipatory Gaze Shifts during Navigation in a Naturalistic Virtual Environment. ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play. London, England, (2015).
5. Blascovich, J., Loomis, J., Beal, A., Swinth, K. R., Crystal, H. L., & Bailenson, J.: Immersive virtual environment technology as a methodological tool for social psychology. Psychological Inquiry, 13(2), 103-124 (2002).
6. Burgoon, J. K., Dunbar, N. E.: An interactionist perspective on dominance-submission: Interpersonal dominance as a dynamic, situationally contingent social skill. Communication Monographs, 67, 96-121 (2000).
7. Canossa, A., Badler, J., Seif El-Nasr, M., Tignor, S., Colvin, R.: In Your Face(t) Impact of Personality and Context on Gameplay Behavior. Foundations of Digital Games. Pacific Grove, CA, (2015).
8. Charney, D. L.: True Psychology of the Insider Spy. Intelligencer: Journal of the US Intelligence Studies. 18.1, 47-54 (2010).
9. 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. Communication Research, 41 (6), 852-876 (2014).
10. 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. SBP-BRiMS 2016. Lecture Notes in Computer Science, 9708. Springer, Cham (2016).
11. Ho, S. M., Warkentin, M.: Leader's dilemma game: An experimental design for cyber insider threat research. Information Systems Frontiers, 19 (2), 377–396 2(5), (2017).
12. IARPA SCITE program (https://www.iarpa.gov/index.php/research-programs/scite/scite-baa) (2015)
13. MacDowell, K. A., Mandler, G.: Constructions of emotion: Discrepancy, arousal, and mood. Motivation and Emotion, 13 (2), 105-124 (1989).
14. Sasaki, T.: A Framework for Detecting Insider Threats using Psychological Triggers. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 3 (1/2), 99-119, (2012).
15. Seif El-Nasr, M., Nguyen, T., Carstensdottir, E., Gray, M., Isaacowitz, D., & Desteno, D.: Social Gaming as an Experimental Platform. Social Believability in Games Workshop at Foundations of Digital Games. (2014).
16. Seif El-Nasr, M., Yan, S.: Visual Attention in 3D Games. International Conference on Advances in Computer Entertainment Technology (ACE), 22-26 (2006).
17. Spitzner, L.: Honeypots: Catching the Insider Threat. In: Proceeding of the 19th Annual Computer Security Applications Conference (ACSAC '03), pp. 170-180. IEEE Computer Society, Washington, DC, USA (2003).