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

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

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

Abstract. Detecting spies' behaviors in virtual environments is an important topic of interest to many organizations and has spurred research for years. Here we address this problem by testing emotional active indicators, a theoretical framework about spies' mental states that may affect their 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 automatic active indicators in the game. Then we used statistical and machine learning techniques to develop models based on the emotional active indicators 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). We also looked at the influence of demographics, personality and other factors on players' choice to betray and their behaviors. While the active indicators used by the statistical and machine learning techniques did not clearly discriminate betrayers from decliners and controls, 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. We thus present this work and the models we developed as new contribution alluding to the use of games as a method to investigate and deeply examine deceptive behavior in a controllable manner, and showing results that reinforce arguments made in previous work.

Keywords: Behavior Modeling, Deceptive Communication, Online games.

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 approach for detecting spies is based on placing 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 espi-

onage/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.

In this paper, we discuss work we embarked in to further explore this theory and approach. In particular, our goal was to design and test a set of automated stimuli, which we call "Active Indicators" (abbr. AIs), and apply them within a simulated controlled online environment (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.

We chose to embed the AIs in an online game, because game environments have proven well suited to study many aspects of human behavior, including the types of behavioral indicators of emotion that are the basis for our AIs (see the Related Work section), and because we expected that the emotional impacts of betrayal could develop relatively quickly against a relatively simple background task offered by a 1-hour long game.

Our work was driven by the following research question: can we use statistical analyses and machine learning techniques to build a behavioral model of betrayal, based on a set of detector signals for the AIs, that can distinguish the behaviors of betrayers (people who chose to betray) from those of non-betrayers (people who chose not to betray) and controls (people who were not given a choice to betray)? This paper describes the game we created, the experimental hypotheses we had about the AIs, the results of our attempts at building such behavioral model, and the unexpected similarities between betrayers' and deceivers' communicative strategies.

2 Related Work

Virtual environments and games have been recommended as a methodological tool to study social psychology [Blascovich et al. 2002], trust [Seif El-Nasr et al. 2014], and learning [Gee 2007]. In the field of economics, there has been much work on using games to study economic decision-making (e.g. [Sanfey et al. 2003], [Camerer and Fehr 2002]). Psychology researchers have used games 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], [Greenfield et al. 1994], [Badler and Canossa 2015], interpersonal space [Bailenson et al. 2003], emotions and motivation [MacDowell and Mandler, 1989], [Adbuhamdeh et al. 2015], personality [Canossa et al., 2015].

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 communicative 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 cues related to time-lag, social and negation in text have the ability to 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 hypothesized that betrayers (people who were asked and agreed to share information with a competitive team, thus betraying their own team) would have exhibited less identification and trust with their teammates, and less focus and diligence on their task, compared to controls and decliners (people who declined to share information with a competitive team). Based on previous work and Subject Matter Experts involved in developing the experimental stimuli with our team, we developed a set of target behaviors for 18 AIs (or stimuli embedded within the game), and we hypothesized that 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:

- **Target Behavior**: "Doesn't devote full attention to job, as a result of anxiety about discovery"; **AI stimulus**: "Al-23: Present a time limited window to complete routine work"; **Detector Signal**: More likely to miss responding
- **Target Behavior**: "Doesn't devote full attention to job, as a result of anxiety about discovery"; **AI stimulus**: "Al-27: Present opportunities to disengage from task"; **Detector Signal**: More likely to disengage

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. Each team earns points when members correctly answer questions about their target mystery stranger. Teammates communicate through text chat to share their theories and help each other answer questions correctly. The team consisted of 1 human player and 3 bots (their automated nature was not disclosed to the human player). To maintain experimental control and comparability across teams, the bot teammate text chat was completely pre-scripted and non-reactive to any of the participant's text chat or other actions.

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.

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 on the Amazon Mechanical Turk crowdsourcing platform (subjects got a $5 reward plus a $2 bonus).

Control group participants played the game with no opportunity to betray their team, while inducement group participants were offered a chance to receive a $2 bonus payment in return for turning over key information to the opponent team. The betrayal inducement was presented to participants at the end of the first round of the game. 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, and a personality survey several days after gameplay. After a few days, participants were asked again to participate in a survey to collect their personality data for an extra $2. We used the NEO-PI and the NPI (Narcissist Personality Inventory).

We ran 2 cycles of the experiment: in Cycle 2, we emphasized both (a) 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 (b) 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. Here we focus on the results of Cycle 2.

4 Results and discussion

We had 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. 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.

We computed correlations between AIs, 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 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 measures are discriminative compared to individual characteristics or self-reported feelings.

The most discriminative active indicator meeting a criteria of > 30% true positive rate (TP) and > 5 true positive to false positive ratio was: teammate chat and the opportunity to respond. Specifically: (a) *betrayers were more likely to engage in a high volume of chat*. The next most discriminative active indicator was the opportunity to rate team and teammate trustworthiness: (b) *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).

Even though the betrayers of our experiments were not requested to actively engage in deceptive communication, they may have used communication strategies analogous to those of deceivers, 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 persuade teammates of being genuine team members rather than spies ("persuasive deception", see [Dunbar et al. 2014], and produced effects are similar to those found by [Anolli et al. 2003] when "lying to a suspicious recipient".

This paper makes two concrete contributions: 1) methodological: the paper shows the use of games as a method to deeply analyze betrayal and deception like behaviors, and 2) the paper presents results, that confirm previous work, showing that betrayers engage in more chat and are more likely to rate their teammates as less trustworthy. 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. Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M.: Interpersonal distance in immersive virtual environments. Personality and Social Psychology Bulletin, 29 (7), 819-833 (2003).
5. 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).
6. 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).

7. 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).
8. 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).
9. Charney, D. L.: True Psychology of the Insider Spy. Intelligencer: Journal of the US Intelligence Studies. 18.1, 47-54 (2010).
10. 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).
11. Gee, J. P.: What Video Games Have to Teach Us About Learning and Literacy. Second Edition. Palgrave Macmillan, New York (2007).
12. Greenfield, P. M., DeWinstanley, P., Kilpatrick, H., & Kaye, D.: Action video games and informal education: Effects on strategies for dividing visual attention. Journal of Applied Developmental Psychology, 15 (1), 105-123 (1994).
13. 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).
14. 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).
15. MacDowell, K. A., Mandler, G.: Constructions of emotion: Discrepancy, arousal, and mood. Motivation and Emotion, 13 (2), 105-124 (1989).
16. Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D.: The neural basis of economic decision-making in the ultimatum game. Science, 300 (5626), 1755-1758 (2003).
17. 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).
18. 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).
19. Seif El-Nasr, M., Yan, S.: Visual Attention in 3D Games. International Conference on Advances in Computer Entertainment Technology (ACE), 22-26 (2006).
20. 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).