Beyond the Prisoner's Dilemma: the Social Dilemmas of Cybersecurity

Jordan Richard Schoenherr[1,2] and Robert Thomson[1]
[1]Army Cyber Institute / Behavioral Science and Leadership Department, US Military Academy
[2]Department of Psychology / Institute of Data Science, Carleton University
Jordan.Schoenherr@carleton.ca; robert.thomson@westpoint.edu

***Abstract*—**The Prisoner's Dilemma represents an ubiquitous approach to security modeling that emphasizes adversarial relationships between actors. Adopting this approach helps understand ambiguous relationships in information domains. Despite the fact that some actors might adopt these frames, the Prisoner's Dilemma reflects only one of many possible social dilemmas. In this paper, we outline a computational approach to cybersecurity based on Interdependence Theory. Interdependence Theory provides a means to decompose pay-off matrices into social influence components based on the amount of control actors and partners have in a situation. It additionally accounts for joint control that develops from the mutual decisions of both players. By focusing on two-person, two-option games, this approach can model many different social situations that reflect nor-mal and anomalous network activity.

***Keywords*—** game theory, Interdependence Theory, cybersecurity, social dilemmas

I. Introduction

Network security is a pervasive and ineliminable aspect of the modern world. Resource limitations within organization requires effective strategies for resources allocation. For instance, Stackelberg competitions [1, 2] model the distributing of limited defensive resources and target selection for defenders and attackers, respectively. Understanding the social-cognitive processes supporting online interactions is crucial. [3, 4, 5] In order to reduce the inherent ambiguity of behavior in information domains, cybersecurity professionals can use these models to develop network defense strategies.

Currently, many cybersecurity analysts implicitly adopt a game-theoretical approach to cyber defense that reflects the Prisoner's Dilemma (PD). [6, 7, 8, 9, 10] The dilemma emerges by providing players with payoffs that incentivize individual defection (i.e., failure to cooperate) despite the possibility of mutual gains. However, the Prisoner's Dilemma reflects only one of many social dilemmas. [11, 12, 13, 14, 15] While, by their very nature, cyberattack and cyber defense strategies reflect PDs, many agents in information domains likely do not share these motivations. Moreover, ethical dimensions must be considered (e.g., privacy; [10]). In the following review, we present a general computational approach to social dilemmas that can be used to developed cyberdefence strategies. Extended Interdependence Theory, we assume that multiple social dilemmas exist based on payoff matrices (e.g., Assurance Dilemma, Chicken Dilemma). By decomposing these payoff matrices, a number of motivational components can be identified that specify the control of the actor, their partner, and their joint control. We argue that viewing cybersecurity in terms of social dilemmas provides a more effective and robust approach to understanding otherwise ambiguous information domains confronting network security analysts.

II. Game Theory: Cooperation and Conflict

Game theoretic models applied to security problem generally assume competition. These approaches have been adopted in cybersecurity. [6, 7, 8] Game theory reflects a prescriptive model based on rational choice, i.e., players should do what is in their interest. Players in these games typically have only two options: to cooperate with other player(s) or to defect and pursue a different strategy.

A ubiquitous assumption in the use of Game Theory is that players are placed into competition with one another. For instance, one of the most prominent games is the Prisoner's Dilemma (PD). The PD reflects a two-person, two-option social dilemma wherein the collective interest of two players is placed in competition with the self-interest of each individual player (see Figure 1, Table 1). Game theoretic approaches have been adopted as a means of explaining and predicting social interaction. However, as we will describe below, the PD reflects only one of many social dilemmas defined by pay-off matrices [11, 13]and research has frequently found violations of self-interested behavior. [16]

The prescriptive models described by Game Theory define ideal strategies. For instance, Nash [17] provide solutions to games, referred as Nash equilibria. A Nash equilibrium reflects a steady state within a game [18]. Once an equilibrium is reached, no player will prefer to change their strategy because doing so would lower their individual payoff. In the PD, the Nash equilibrium in is for both players to defect. This stems from the structure of pay-off matrix such that it is in each individual's interest to defect given the greater pay-off [17].

Despite this prediction, early studies of the PD demonstrated violations of this individual rational strategy. [16] Violations of these equilibria can be accounted for in a number of ways: 1) normative models of rationality do not apply to humans in whole or in part, 2) normative models of rationality might include collective rationality that supports cooperation, 3) the pay-off matrix might not be accessible to players in a given round of play, and/or 4) despite the availability of a pay-off matrix, players might believe another schema is more applicable.

A. The Description-Experience Gap

A prominent means to understand these discrepancies between rational prescriptive behavior and observed performance is the description-experience gap (DEG; [19]). The DEG assumes that there are two broad sources of information. The **description** component assumes that players' expectations will be determined by how the game is presented

to players. For instance, a game might be described as cooperative or competitive or a pay-off matrix might promote competition as opposed to cooperation (i.e., the PD).

In contrast, the **experience** components assume that players learn through the course of the game. Rather than expectations being fixed, players adjust them throughout the course of the study. [20] [21] For instance, in the context of an Iterative Prisoner's Dilemma (IPD), while a pay-off matrix might set a player's expectations during the first round of play, the behavior of a player might cause them to adjust their strategy.

The gap occurs when a player's expectations are violated because of the actions of other social agents within the game. Experimentally, this can be examined by providing some players with a frame while failing to providing others players with a frame. In most studies examining frames, a frame is either invoked or not. However, there are likely different levels of information available. For instance, Martin et al. [22] used an IPD – they varied the extent to which players had information concerning other's actions within the game. Participants had no information (no info), basic knowledge of their interdependence in the task (min info), specific feedback concerning the actions or others (mid info), or information concerning the overall payoffs of performing actions within a task (max info). They found that moderate- to [high-levels of information produced more cooperative behavior relative to low-levels or no information.

B. Cooperative Motivation

Humans are an ultra-social species given the requirements of coordinating complex interdependent activities. Whether explicitly defined or implicitly learned, systems of normative regulation emerge over time that impact our cooperative behavior. For instance, Tenbrunsel and Messick [23] examined the use of sanctioning systems. Sanctioning systems are believed to have a direct effect wherein payoff structures increase and decrease specific behaviors and an indirect effect by changing the expectations of social agents within the system. [24, 25, 26] They found that participants that perceived the task as a business decision when strong sanctions were referenced were more likely to engage in a calculus than those that perceived the task as an ethical decision when no sanctions or only weak sanctions were referenced.

In another early study to examine this in PD, Liberman, Samuels, and Ross [27] provided participants with an iterative prisoner's dilemma (IPD) wherein the game were referred to as the "Wall Street Game" or the "Community Game". They assumed that while referring to the game as the "Wall Street Game" would promote low levels of cooperation comparable to typical PD tasks, referring to the game as the "Community Game" would promote more cooperation (for related results, see [28, 29]). Similarly, these frames can become entrenched within a group over time, leading it to become a facet of a group's cultural traditions. For instance, Henrich, Heine, Norenzayan [30] suggested that there are two kinds of group exchange patterns that have emerged based on market integration. In a highly integrated market, social agents are interdependent. Consequently, they are more inclined to expect adherence to fairness norms and punish defectors even when they must incur a cost.

C. Individual Differences

Another prominent source of variation in these games are the values that players bring into the games. Some players might be more inclined to collaborate than others. De Dreu and McCusker [31] first assessed their participants' social value orientation (SVO) and identified those with high levels of prosocial values ("prosocials") and those participants who were individualists. Participants were then provided with a social dilemma framed in terms of either gains or losses. They found that cooperation rates were a function of SVO and the loss framing. They found that prosocials were more likely to cooperate in a loss frame than in a gain frame. In contrast, individualists were more likely to cooperate in a gain frame than in a loss frame. These results suggest that both frames and individual differences affect performance in these tasks.

II. Relational Structures and Interdependence Theory

Physical and social environments represent ambiguous structures to social agents. When they are compatible, experiences can prove to be an effective guide to navigating a social situation. In the absence of prior experience, we can use schemata acquired through direct experience with similar situations or indirectly through processes of cultural transmission. Recent studies have found some promising evidence that framing can affect players' responses when playing a security dilemma. [29] Consequently, much like DEG, payoff matrices and other social structures retained in memory can compete.

A. Pay-Off Matrices and the Structure of Social Situations

Inter-situation and cross-cultural differences might obscure the deep-structure of social situations. For instance, Interdependence Theory [32], [33] reflects an attempt to describe all possible social situations in terms variance components that can be compared. One such broad attempt has resulted in an *Atlas of Interpersonal Situations* [34] which demonstrates that any situation can be represented in pay-off matrices that can be understood in terms of variance components.

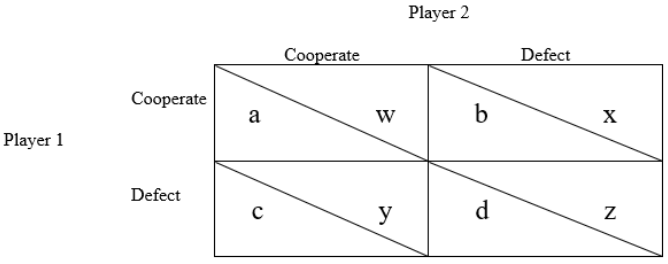

Fig. 1. General structure of a two-person, two-option social dilemmas.

The simplest social dilemma reflects a two-person, two-option social dilemma. These social dilemmas have a basic structure (**Fig. 1**).

In addition to the PD, a number of other social dilemmas can be described in terms payoff matrices. Consider two prominent alternatives that reflect distinct social dilemmas [12], [35] and motivate different behavior when presented to players

[36]: the Assurance Dilemma and the Chicken Dilemma (Table 1). In the Assurance Dilemma, individual and collective outcomes are maximized when both parties cooperate. For instance, two companies benefit collectively and individually if they both upgrade malware programs (e.g., Example 2, Figure 2). In the Chicken Dilemma individual reward is greatest when another player defects, while collective reward is greatest when both players defect. Adversarial cyber operations between two nation-states would reflect such a pattern wherein cessation of network intrusions would benefit both actors while the cessation of one nation state's operations would benefit another if they failed to cease their cyber operations (e.g., Example 3, Figure 2 to cease their cyber operations (e.g., Example 3, Figure 2).

Table 1. Payoff structure and types social dilemmas.

Type of Social Dilemma	Player 1	Player 2
Prisoner's Dilemma	c > a > d > b	x > w > z > y
Assurance Dilemma	a > c = d > b	w > x ≥ z > y
Chicken Dilemma	b > d > c > a	y > z > x > w

An important observation in subsequent extensions in Interdependence Theory is that players can *transform* pay-off matrix. Corresponding to a similar logic of DEG, while a player might be presented with one payoff matrix (e.g., PD) they might instead behave in a manner that reflects another payoff matrix (e.g., Assurance Dilemma). Thus, when the DEG is observed, Interdependence Theory suggest that responses can be used to infer what social dilemma players *believe* they are being presented.

B. Control Variance Components

Game Theory adopts (an essentially) deterministic approach to the pay-off matrix: when pay-offs are provided, players motivations will change. However, this does not necessarily consider what factors are changing a player's motivations. Another theoretical means to consider social dilemmas is in terms of the extent to which pay-offs for decisions by each player translate into situational control.

Lewin [37], [38] drew an analogy with physics, suggesting that social situations were defined by a 'field of forces.' Approaching a pay-off matrix in these terms, the decisions of each player exert control over the situation. Consequently, a payoff matrix can be decomposed into variance components that determine situational control.

For illustration purposes, imagine two companies (C1 and C2) that have networks that are vulnerable to attackers. Attackers have limited resources and will therefore only have time to attack a company that has the weakest network defenses. We can construct two payoff matrices which correspond to a PD (**Fig 2. a**). In terms of Interdependence Theory [32], [33], there are three main variance components: actor control (AC), partner control (PC), and joint control (JC). Together, AC, PC, and JC create separate matrices (ACM, PCM, and JCM) which reflect the individual control components that influence a social agent within a given situation.

Actor Control. For **Actor Control.** AC is defined as the amount of control a given player has over their outcomes. For instance, if C1 decides to upgrade its firewalls beyond those of C2, AC reflects the effect of C1 on C2's decision. It reflects a difference score that reflects the outcomes if C1 acts or if C1 does not act.

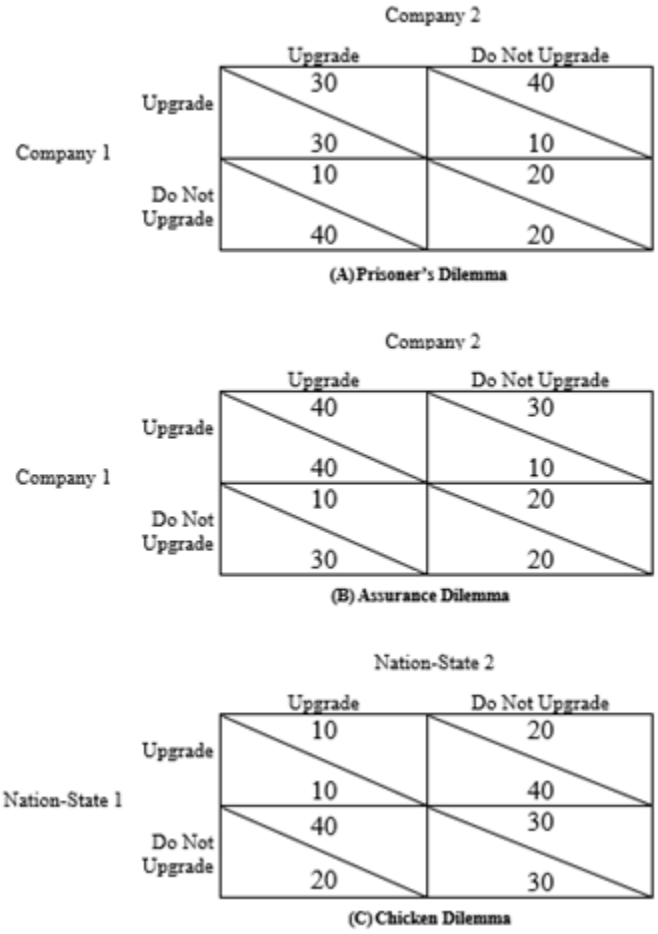

Figure 2. Examples of Three Representative Social Dilemmas.

In order to compute AC, the payoff matrix is decomposed by considering the outcomes of C1s decision to upgrade relative to if it does not decide to upgrade. First, we obtain the average outcome when C1 upgrades its network ([30 + 10] / 2 = 20) and the average outcome if it does not upgrade its network ([40 + 20] / 2 = 30). Then, we obtain a difference score between when they do not upgrade compared to when they upgrade (30 − 20 = 10). When C2 values are computed, they jointly create the ACM. For instance, given $AC_{C1} = 10$, AC's decision to upgrade or not upgrade exerts 10 units of control over their own outcomes within this social dilemma.

Partner Control. PC reflects the amount of control that the competing company, C2, has over C1. In a like manner, it reflects a difference score that reflects the outcomes if C2 acts or if C2 does not act. First, we obtain the average outcome for C1 when C2 upgrades its network ([30 + 40] / 2 = 35) and the average outcome for C1 if C2 does not upgrade its network ([10 + 20] / 2 = 15). Then, we obtain a difference score (35 − 15 = 20). Values for the JCM are derived in the same manner as those of the PCM.

Joint Control. Finally, JC reflects the amount of control that the joint decisions of both companies have on C1's outcomes. Unlike AC and PC, deriving JC requires the use of both ACM and PCM. This requirement is a consequence of the outcomes that occur at the level of specific cells within the payoff matrix. In order to compute a cell of the JCM, a value of a cell in the payoff matrix is identified (e.g., the outcome for C1 if both C1 and C2 upgrade). Then, the sum of the corresponding values in the ACM (i.e., 0) and PCM (i.e., 20) is subtracted from the value in the JCM. Thus, the values in the JCM reflect the remaining variance that is not attributable to AC or PC. The value of JC can then be obtained such that the difference for JCM for C1 ([35+35]/2 = 35) and for C2 ([35+35]/2 = 35), is equal to 0.

Importantly, AC, PC, and JC can be used to understand what social influences will be experienced by social agents. For instance, given that AC < PC in the above example, it suggests that C2 has more control over the outcomes of this situation than C1. Alternatively, if C1 or C2 violate predictions based on the values of AC and PC, it might suggest that they do no perceive the social affordances of the situation.

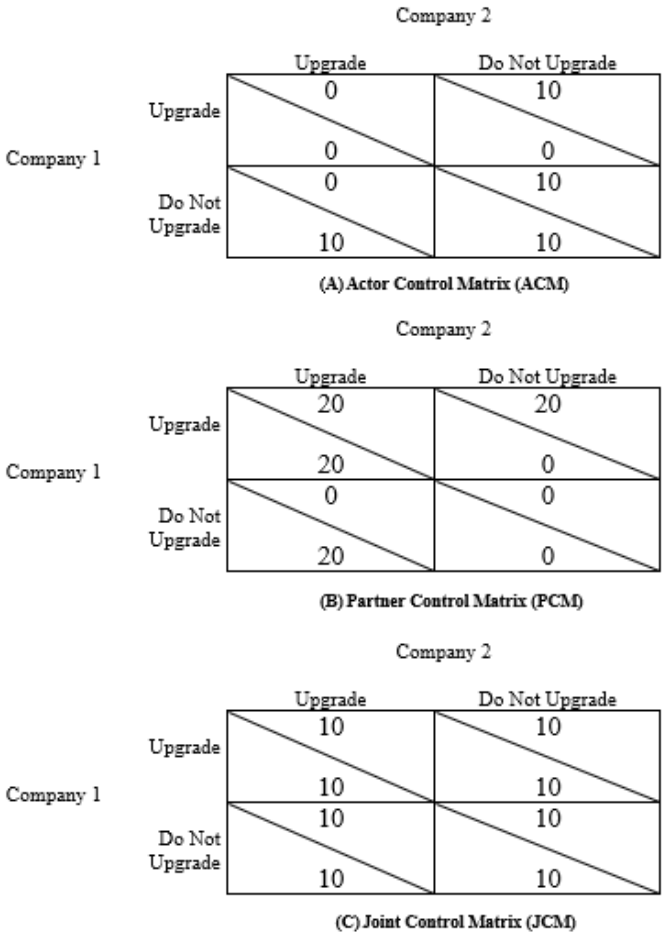

Figure 3. Examples for ACM, PCM, and JCM.

Similarly, the AC, PC, and JC for the Assurance Dilemma are provided in Figure 2. In this case, given the payoff matrix symmetry and that both C1 and C2 have the highest payoff for joint action (mutual upgrading or mutual failure to upgrade), the JCM is equal to the payoff matrix: i.e., the social dilemma if the AD provided here is defined by mutual interdependence.

C. Higher-Order Variance Components

Interdependence Theory [32] has the additional benefit of assuming that composite motivational factors can be derived from a recombination of AC, PC, and JC variance components. While an exhaustive discussion is outside the scope of the present study, a few higher-order variance components are worth noting. For instance, the *degree of interdependence* (DI) can be calculated by considering the sum of the variance associated with control that other players have over us (i.e., PC and JC) relative to the variance associated with all forms of control (i.e., AC, PC, and JC). For any given player, DI is given by:

$$DI = \frac{\mathrm{PC}^2 + \mathrm{JC}^2}{\mathrm{AC}^2 + \mathrm{PC}^2 + \mathrm{JC}^2}$$

Large values of DI are obtained when AC is small. Conceptually, when DI is large there is a high degree of interdependence in a given social situation. When DI is small, an individual social agent has considerable control over the situation. For instance, in our example of PD, the interdependence for Company 1 (C1) would reflect the contributions of PC and JC (i.e., $2^2 + 0^2 = 4$) relative to total control (i.e., $1^2 + 2^2 + 0^2 = 5$), or 0.8 suggesting that C1 is highly dependent on the actions of C2 and their joint actions.

Asymmetries in dependencies can be obtained in a similar manner. *Asymmetric dependency* (AD) is simply computed by obtaining the difference between the dependence of on social agent and that of another. In our example,

$$\mathrm{AD} = \mathrm{DI}_{C1} - \mathrm{DI}_{C2}$$

In simple terms, this function serves to check the symmetry of the payoff matrices. Given that the pay-off matrix is symmetric for C1 and C2 in our example, no asymmetric dependencies exist. Consequently, in a similar manner to the interpretation of the individual control components, we can conclude that while both C1 and C2 are significantly influenced by the actions of their opponent, they exert equal social influence in this scenario.

Again, a similar calculation can be performed for the Assurance Dilemma. In this case, DI the contributions of PC and JC (i.e., $20^2 + 10^2 = 500$) relative to total control (i.e., $0^2 + 20^2 + 10^2 = 500$), give AD = 1.

Due to this summative ability, Interdependence Theory provides a principled means to describe the social forces within a situation. It additional provides a rich source of predictions at a fine-grained level that an examination of social dilemmas as a whole does not necessarily address. However, whether the variance components can be distinguished by players or have motivational primacy, requires further theoretical and empirical investigation. For instance, like the Covariation Model of Attribution also developed by Kelley [39, 40, 41], claims that assume humans consider the variance components of a situation to determine whether an internal (agent-based) or external (situation-based) attributions are appropriate, humans likely do not assess control and dependency in this manner. Even early

studies demonstrated that attributions are not primarily dependent on these processes. [42, 43, 44] Similarly, humans are not naïve scientists that weigh evidence as the model implies. Consequently, verification that the basic variance (AC, PC, and JC) or their use in higher-order calculations (e.g., DI and AD) requires further work. Nevertheless, this comprehensive theory sets out numerous empirical predictions that can be tested and modelled.

III. Applications in Information Domains

Despite the theoretical and empirical evidence that many motivation schemata are available to players in security games, translating the evidence from physical domains to information domains requires a consideration of features of online interactions. An essential feature of social interaction is that one feels that they are being monitored. Indeed, this is an essential feature of any social monitoring and regulation system [24]. The experience of being evaluated can alter participants' responses. Indeed, reminders of state- or religion-based social institutions results in a greater adherence to fairness norms [45].

Despite the requirement of user names, social media promotes anonymity. Archival studies [46] and ecological experiments [47] have demonstrated that the experience of anonymity in terms of being in a large group and/or not being identifiable results in more antisocial behaviour (cf. [48]). Studies of online behavior produce similar results [49, 5], with evidence also suggesting that more extreme attitude changes occur in online environments [50].

Professionals in information and computer science must also consider how these models apply to human interactions with autonomous and intelligent systems. Namely, the realism of bot behavior is important for trust, cooperation, and sustained interaction. [51, 52, 53, 54] Using an IPD, Ishowo-Olko et al. [55] paired users with human or bots. They found that when users believed a bot was human, they were more likely to cooperate with the bots then when the player was aware of the bots identity. However, earlier studies suggest that while humans are just as likely to reciprocate with nonhuman agents, they are less likely to cooperate with them. [56] While Interdependent Theory is a prescriptive and predictive model of human behaviour, it can be extended to bots as a means to simulate behavioral interaction as well as a means to use these norms to detect violations of these norms.

IV. Conclusions

Advances in social psychological research provide can provide considerable insight into the otherwise ambiguous and ill-define domain of online interactions. However, we must be cautious when adopting theoretical models that consider only a subset of social interactions. The introduction of Game Theory [17] provided a useful tool for understand how the structure of social situations can be modelled, including the notion of equilibria that can be determined by means of assuming the pursuit of rational self-interest. However, violations of rational self-interest suggest that players either fail to adhere to these principles or are utilizing other principles. [16]

Interdependence Theory [32], [33] offers a more comprehensive means to address social dilemma. Consequently, the excessive focus on PD neglects other kinds of social dilemma structures that might inform players' performance (e.g., AD or CD). Second, players might neglect the pay-off matrix to focus on an overall frame. These frames can either be primed by instructions, reflect individual differences, or have been learned through repeated exposure. Finally, we assume that social factors such as the experience of anonymity and the perceived connection with a group (i.e., a shared social identity) will also be crucial in determine when and what norms are used. In their simplest form, this might decrease and increase the adherence to norms that support individual- and collective-interests, respectively.

In general, we do not assume that everyone within cyberspace perceives themselves as an 'attacker' or a 'defender'. Rather, the social frames of information domains likely vary from a communal space for sharing information to the perception of anonymity and ambiguity. This can be partially attributed to the open-ended nature of information and network systems that underpin this domain coupled with a failure to read the terms of service – a pay-off matrix – for an application. In short, assuming that cybersecurity is an issue for an individual social agent requires that they understand the dual-use nature of the technology, for instance, that they can be both a defender and an attack vector. Even if users adopt such a frame, additional ethical considerations can also be salient (e.g., privacy; [10]). More effective understanding and communicating these norms requires a greater incorporation of existing knowledge from the behavioral and social sciences.

Acknowledgements

Research was sponsored by the Army Research Laboratory and was accomplished under the Cooperative Agreement Number W911NF-19-2-0223. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for the Government purposed notwithstanding any copyright notation herein.

References

[1] H. von Stackelberg, Marktform undGleichgewicht., Springer, 1934.

[2] Z. Yin, D. Korzhyk, C. Kiekintveld, V. Contizer and M. Tambe, "Stackelberg vs. Nash in Security Games: interchangability, Equivalence, and Uniqueness.," in *in Proceedings of the Internaional Conference on Autonomous Agents and Multiagent Systems (AAMAS).*, 2010.

[3] G. Riva and C. Galimberti, "The psychology of cyberspace: A socio-cognitive framework to computer-mediated communication," *New ideas in Psychology,* vol. 15, pp. 141-158, 1997.

[4] G. Riva, "The sociocognitive psychology of computer-mediated communication: The present and future of technology-based interactions," *Cyberpsychology & Behavior,* vol. 5, pp. 581-598, 2002.

[5] A. G. Zimmerman and G. J. Ybarra, "Online aggression: The influences of anonymity and social modeling," *Psychology of Popular Media Culture,* vol. 5, pp. 181-, 2016.

[6] A. Sokri, "Game Theory and Cyber Defense," in *Games in Management Science*, Cham, Springer, 2020, pp. 335-352.

[7] N. Kostyuk, "The Digital Prisoner's Dilemma: Challenges and Opportunities for Cooperation," in *IEEE CyberSummit*, 2013.

[8] A. E. Chukwudi, E. Udoka and I. Charles, "Game Theory Basics and Its Application in Cyber Security," *Advances in Wireless Communications and Networks,* vol. 3, pp. 45-49, 2017.

[9] B. Buchanan, The cybersecurity dilemma: Hacking, trust, and fear between nations, Oxford University Press, 2016.

[10] M. D. Cavelty, "Breaking the cyber-security dilemma: Aligning security needs and removing vulnerabilities," *Science and Engineering Ethics,* vol. 20, pp. 701-715, 2014.

[11] H. H. Kelley and J. W. Thibaut, Interpersonal Relations: A Theory of Interdependence, New York: John Wiley, 1978.

[12] A. Rapoport and M. Guyer, "A taxonomy of 2 × 2 games.," *General Systems,* vol. 11, pp. 203-214, 1966.

[13] D. Balliet, J. M. Tybur and P. A. M. Van Lange, "Functional Interdependence Theory: An Evolutionary Account of Social Situations," *Personality and Social Psychology Review,* pp. 1-28, 2016.

[14] E. van Dijk and H. Wilke, "Decision-induced focusing in social dilemmas: Give-some, keep-some, take-some, and leave-some dilemmas," *Journal of Personality and Social Psychology,* vol. 78, pp. 92-104, 2000.

[15] S. S. Komorita and C. D. Parks, Social Dilemmas, Boulder: Westview, 1996.

[16] A. C. A. Rapoport, Prisoner's Dilemma: A Study in Conflict and Cooperation., Ann Arbor: University of Michigan Press, 1965.

[17] J. NASH, "Equilibrium points in n-person games.," in *Proceedings of the National Academy of Sciences*, 1950, pp. 48-49.

[18] Y. Z, K. D, C. Kiekintveld, V. Contizer and M. Tambe, "Stackelberg vs. Nash in Security Games: interchangability, Equivalence, and Uniqueness.," in *in Proceedings of the Internaional Conference on Autonomous Agents and Multiagent Systems (AAMAS).*, 2010.

[19] D. U. Wulff, M. Mergenthaler-Canseco and R. Hertwig, "A meta-analytic review of two modes of learning and the description-experience gap," *Psychological Bulletin,* vol. 144, pp. 140-176, 2018.

[20] R. D. &. S. P. Luce, "Preference, utility, and subjective probability.," in *In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), Handbook of mathematical psychology, Vol. 3*, New York, Wiley, 1965, p. 249–410.

[21] W. Lee, Decision theory and human behavior, New York: Wiley, 1971.

[22] J. M. Martin, C. Gonzalez, I. Juvina and C. Lebiere, "A Description–Experience Gap in Social Interactions: Information about Interdependence and Its Effects on Cooperation," *Journal of Behavioral Decision Making,* vol. 27, p. 349–362, 2014.

[23] A. E. Tenbrunsel and D. M. Messick, " Sanctioning systems, decision frames, and cooperation," *Administrative Science Quarterly,* vol. 44, pp. 684-707, 1999.

[24] T. Yamagjshi, "The Provision of a Sanctioning System as a Public Good," *Journal orpersonality and Social Psychology,* vol. 51, pp. 110-116, 1986.

[25] T. Yamagishi, "Seriousness of Social Dilemmas and the Provision of a Sanctioning System," *Social Psychology Quarterly,* vol. 51, pp. 32-42, 1988.

[26] T. Yamagishi, "Group Size and the Provision of a Sanctioning System in a Social Dilemma," in *A Social Psychological Approach to Social Dilemmas*, Oxford, Pergamon, 1992, p. 267–287 .

[27] V. Liberman, S. M. Samuels and L. Ross, "The name of the game: Predictive Reputations Versus Situational Labels in Determining Prisoner's Dilemma Game Moves," *Personality and Social Psychology Bulletin,* vol. 30 , pp. 1175-1185, 2004.

[28] M. M. Pillutla and X.-P. Chen, "Social Norms and Cooperation in Social Dilemmas: The Effects of Context and Feedback," *Organizational Behavior and Human Decision Processes,* vol. 78, pp. 81-103, 1999.

[29] M. Grinberg, E. Hristova and M. Borisova, "Cooperation in Prisoner's Dilemma Game: Influence of Social Relations," in *Proceedings of the Cognitive Science Society*, 2012.

[30] J. Henrich, S. J. Heine and A. Norenzayan, "The weirdest people in the world?," *Behavioral and Brain Sciences,* vol. 33, pp. 61-83, 2010.

[31] C. K. De Dreu and C. McCusker, "Gain–loss frames and cooperation in two-person social dilemmas: a transformational analysis," *Journal Personality and Social Psychology,* vol. 72, p. 1093–1106, 1997.

[32] H. H. Kelley and J. W. Thibaut, Interpersonal relations: A theory of interdependence, New York: John Wiley, 1978.

[33] D. Balliet, J. M. Tybur and P. A. M. Van Lange, "Functional Interdependence Theory: An Evolutionary Account of Social Situations," *Personality and Social Psychology Review,* pp. 1-28, 2016.

[34] H. H. Kelley, J. G. Holmes, N. Kerr, H. Reis, C. Rusbult and P. A. Van Lange, An Atlas of Interpersonal Situations, Cambridge: Cambridge University Press, 2003.

[35] D. M. Messick, "Alternative logics for decision making in social settings," *Journal of Economic Behavior & Organization,* vol. 39, pp. 11-28, 1999.

[36] N. Halevy, E. Y. Chou and K. Murnighan, "Mind games: The mental representation of conflict," *Journal of*

Personality and Social Psychology, vol. 102, pp. 132-148, 2012.

[37] K. Lewin, "Field Theory and Experiment in Social Psychology," *American Journal of Sociology,* vol. 44, p. 868–896, 1939.

[38] B. Burnes and B. Cooke, "Kurt Lewin's Field Theory: A Review and Re-evaluation," *International Journal of Management Reviews,* vol. 15, pp. 408-425, 2013.

[39] H. H. Kelley, "Attribution theory in social psychology," in *In D. Levine (Ed.), Nebraska Symposium on Motivation*, Lincoln, University of Nebraska Press, 1967.

[40] H. H. Kelley, "The process of causal attribution," *American Psychologist,* vol. 28, pp. 107-128, 1973.

[41] H. H. Kelley and J. L. Michela, "Attribution theory and research.," *Annual Review of Psychology,* vol. 31, pp. 457-501, 1980.

[42] G. J. O. Fletcher, "The analysis of verbal explanations for marital separation: Implications for attribution theory," *Journal of Applied Social Psychology,* vol. 13, pp. 245-258, 1983.

[43] P. T. Lewis, " A naturalistic test of two fundamental propositions: Correspondence bias and the actor-observer hypothesis," *Journal of Personality,* vol. 63, pp. 87-111, 1995.

[44] M. W. Passer, H. H. Kelley and J. L. Michela, "Multidimensional scaling of the causes for negative interpersonal behavior," *Journal of Personality and Social Psychology,* vol. 36, pp. 951-962, 1978.

[45] A. F. Shariff and A. Norenzayan, "God is watching you: Priming god concepts increases prosocial behavior in an anonymous economic game.," *Psychological Science,* vol. 18, pp. 803-809, 2007.

[46] L. Mann, "The baiting crowd in episodes of threatened suicide," *Journal of Personality and Social Psychology,* vol. 41, pp. 703-709, 1981.

[47] E. Diener, S. Fraser, A. L. Beaman and R. T. Kelem, "Effects of deindividuating variables on stealing by Halloween trick-or-treaters.," *Journal of Personality and Social Psychology,* vol. 33, pp. 178-183, 1976.

[48] C. M. Smith, P. Dzik and E. Fornicola, " Threatened suicide and baiting crowd formation: a replication and extension of Mann (1981)," *Social Influence,* p. DOI: 10.1080/15534510.2019.1669488, 2019.

[49] J. G. Phillips and L. Mann, "Suicide baiting in the internet era," *Computers in Human Behavior,* vol. 92, pp. 29-36, 2019.

[50] C. Sia, B. C. Y. Tan and K. Wei, "Group polarization and computer-mediated communications: effects of communication cues, social presence, and anonymity," *Information Systems Research,* vol. 13, p. 70–90, 2002.

[51] J. Berkeley, J. P. S. Dietvorst and C. Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology,* vol. 144, p. 114–126, 2015.

[52] S. Kiesler, L. Sproull and J. Miller, "A prisoner's dilemma experiment on cooperation with people and human-like computers," *Journal of Personality and Social Psychology ,* vol. 70, p. 47–65, 1996.

[53] T. Merritt and K. McGee, "Protecting artificial team-mates: more seems like less," in *Proceedings. SIGCHI Conference on Human Factors in Compu-ting Systems*, 2012, pp. 2793-2802.

[54] M. Oudah, V. Babushkin, T. Chenlinangjia and J. W. Crandall, "Learning to interact with a human partner," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, p. 311–318.

[55] F. Ishowo-Oloko, J. F. Bonnefon, Z. Soroye, J. Crandall, I. Rahwan and T. Rahwan, "Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation," *Nature Machine Intelligence,* vol. 1, pp. 517-521, 2019.

[56] E. B. Sandoval, J. Brandstetter, M. Obaid and C. Bartneck, "Reciprocity in Human Robot Interaction – A Quantitative Approach Through The Prisoner's Dilemma And The Ultimatum Game," *International Journal on Social Robotics,* vol. 8, pp. 303-317, 2015.