International Journal of Advanced Smart Convergence Vol.13 No.3 41-47 (2024)
http://dx.doi.org/10.7236/IJASC.2024.13.3.41

IJASC 24-3-5

Bayesian Game Theoretic Model for Evasive AI Malware Detection in IoT

Jun-Won Ho

Professor, Division of Information Security, Seoul Women's University, South Korea
jwho@swu.ac.kr

Abstract

In this paper, we deal with a game theoretic problem to explore interactions between evasive Artificial Intelligence (AI) malware and detectors in Internet of Things (IoT). Evasive AI malware is defined as malware having capability of eluding detection by exploiting artificial intelligence such as machine learning and deep leaning. Detectors are defined as IoT devices participating in detection of evasive AI malware in IoT. They can be separated into two groups such that one group of detectors can be armed with detection capability powered by AI, the other group cannot be armed with it. Evasive AI malware can take three strategies of Non-attack, Non-AI attack, AI attack. To cope with these strategies of evasive AI malware, detector can adopt three strategies of Non-defense, Non-AI defense, AI defense. We formulate a Bayesian game theoretic model with these strategies employed by evasive AI malware and detector. We derive pure strategy Bayesian Nash Equilibria in a single stage game from the formulated Bayesian game theoretic model. Our devised work is useful in the sense that it can be used as a basic game theoretic model for developing AI malware detection schemes.

Keywords: *Bayesian Game Theoretic Model, Evasive AI Malware,*

1. Introduction

Evasive malware can be defined as malware having capability of avoiding the detection. Since Artificial Intelligence (AI) is known to be useful for enhancing the performance of various systems, attacker may be interested in creating evasive AI malware that deploys AI to magnify the evasiveness of malware. To explore this threat that can be occurred in Internet of Things (IoT), we devise a game theoretic model to analyze the interactions between evasive AI malware and detector in IoT, inspiring design of evasive AI malware detection scheme. In particular, we formulate a Bayesian game theoretic model with players of evasive AI malware and detector. We derive pure strategy Bayesian Nash Equilibria in a single stage game.

Manuscript Received: July. 6. 2024 / Revised: July. 12. 2024 / Accepted: July. 18. 2024
Corresponding Author: jwho@swu.ac.kr
Tel: +82-2-970-5607, Fax: +82-2-970-5981
Author's affiliation: Professor, Division of Information Security, Seoul Women's University, South Korea

2. Related Work

As directly relevant work, there is input-driven evasive malware work [2] in which two-player Bayesian signaling game is formulated to analyze the interactions between input provider and input receiver in IoT. Dynamic analyzer and human user are treated as types of input provider and input-driven evasive malware is considered as type of input receiver. Pure strategy and mixed strategy Bayesian Nash equilibria are attained for a single-stage game, weak sequential equilibrium is gained for a multi-stage game. Although Bayesian game theoretic model is used as in [2], our proposed work deals with a new evasive AI malware in IoT with different Bayesian game theoretic problem formulation to [2]. As a result, our pure strategy Nash equilibria is distinct to [2].

[1] explores how Android evasive malware can escape from dynamic analysis. In [5], evasive malware is studied in speculative execution environment. Sandbox evasion scheme is devised in [4]. In [3], bare-metal analysis is harnessed for evasive malware detection.

3. Bayesian Game Theoretic Model for Detection of Evasive AI Malware in IoT

In the sense that evasive AI malware in IoT can exploit AI to evade the detection efficiently and effectively, it is imperative to examine the interactions between evasive AI malware and detector in IoT. To fulfill this need, we employ Bayesian signaling game with two players. Player 1 is defined as evasive AI malware detector. Player 2 is defined as evasive AI malware. In our Bayesian signaling game, player 2 has one type of evasive AI malware and player 1 knows the type of player 2. However, player 2 does not recognize the type of player 1, which has two types of non-AI detector and AI detector. Hence, player 2 needs to determine the type of player 1 in accordance with the strategies taken by player 1. The main reason why player 1 has two types of Non-AI detector and AI detector is because low-end IoT devices acting as Non-AI detectors need to perform malware detection based on non-AI functionality due to their limited resources while high-end IoT devices acting as AI detectors can utilize AI functionality for evasive AI malware detection.

Table 1: Notations used in our game.

Notation	Denotation
C_{na}	Cost incurred by launching Non-AI attack
C_{nd}	Cost incurred by running Non-AI defense
C_{aa}	Cost incurred by launching AI attack
C_{ad}	Cost incurred by running AI defense
I_{aa}	I_{aa}=1 if AI attack succeeds I_{aa}=0 if AI attack fails

Notation	Denotation
I_{na}	I_{na}=1 if Non-AI attack succeeds I_{na}=0 if Non-AI attack fails
G_{aa}	Gain acquired by player 2 when AI attack succeeds
G_{ad}	Gain acquired by player 1 with type of AI detector when to defend AI attack
G_{na}	Gain acquired by player 2 when Non-AI attack succeeds
G_{nd}	Gain acquired by player 1 with type of Non-AI detector when to defend Non-AI attack

In the sense that player 2 generally obtains more gain in AI attack success than non-AI attack success, we can assume that Gaa > Gna holds. From the perspective that it is generally more difficult to detect AI attack than non-AI attack and hence more gain can be attained by defending AI attack than non-AI attack. Thus, we can assume that Gad > Gnd holds. In the sense that more cost is generally incurred in launching AI attack and defense against AI attack than non-AI attack and defense against non-AI attack, we can assume that Caa > Cna, Cad > Cnd. Moreover, in the sense that player 1 with type of non-AI detector (resp. AI detector) can perform Non-AI (resp. AI) defense strategy because he takes more gain than cost incurred by adopting Non-AI (resp. AI) defense strategy, we can assume that Gnd > Cnd (resp. Gad > Cad) holds.

Player 1 with type of Non-AI detector has two strategies of Non-defense and Non-AI defense, player 1 with type of AI detector has two strategies of Non-defense and AI defense. Player 2 with type of evasive AI malware has three strategies of Non-attack, Non-AI-attack, AI attack. Notations used in our game is presented in Table 1. More specifically, notations regarding gains and costs used in our game are defined in Table 1. Table 2 exhibits strategies and payoffs of player 1 with type of Non-AI detector and player 2 with type of evasive AI malware. Table 3 displays strategies and payoffs of player 1 with type of AI detector and player 2 with type of evasive AI malware. More specifically, zero payoff means no benefit/loss and negative (resp. positive) payoff indicates loss (resp. gain) in both Tables 2 and 3.

Table 2: Strategies and payoffs of player 1 with type of Non-AI detector and player 2 with type of evasive AI malware.

		Player 2		
		Non-attack	**Non-AI attack**	**AI attack**
	Non-defense	(0,0)	$(-G_{na}, G_{na}-C_{na})$	$(-G_{aa}, G_{aa}-C_{aa})$

		Player 2		
Player 1	Non-AI defense	$(-C_{nd}, 0)$	$(-I_{na}G_{na}+(1-I_{na})G_{nd}-C_{nd}, I_{na}G_{na}-(1-I_{na})G_{nd}-C_{na})$	$(-G_{aa}-C_{nd}, G_{aa}-C_{aa})$

Table 3: Strategies and payoffs of player 1 with type of AI detector and player 2 with type of evasive AI malware.

		Player 2		
		Non-attack	Non-AI attack	AI attack
Player 1	Non-defense	(0,0)	$(-G_{na}, G_{na}-C_{na})$	$(-G_{aa}, G_{aa}-C_{aa})$
	AI defense	$(-C_{ad}, 0)$	$(G_{ad}-C_{ad}, -G_{ad}-C_{na})$	$(-I_{aa}G_{aa}+(1-I_{aa})G_{ad}-C_{ad}, I_{aa}G_{aa}-(1-I_{aa})G_{ad}-C_{aa})$

Theorem 1. A pure strategy Bayesian Nash equilibrium is ((Non-defense if Non-AI detector, Non-defense if AI detector), AI attack) under the condition that $I_{aa}=1$ and $G_{aa}-C_{aa} > G_{na}-C_{na}$ hold.

Proof. (1) We define E_2(Non-AI) and E_2(AI) as the expected payoff of Non-AI attack and AI-attack strategies of player 2 with type of evasive AI malware when player 1 with type of Non-AI detector sticks to Non-defense strategy and player 1 with type of AI detector sticks to Non-defense strategy.

$$\text{If } I_{aa}=1,\ G_{aa}-C_{aa} > G_{na}-C_{na},$$
$$\text{we have } E_2(AI) = q(G_{aa}-C_{aa}) + (1-q)(G_{aa}-C_{aa}) = G_{aa}-C_{aa}$$
$$E_2(Non-AI) = q(G_{na}-C_{na}) + (1-q)(G_{na}-C_{na}) = G_{na}-C_{na}$$

Because $G_{aa}-C_{aa} > G_{na}-C_{na}$ and $G_{aa} > C_{aa}$, $E_2(AI) > 0$ and $E_2(AI) > E_2(Non-AI)$ hold. As a result, the optimal strategy of player 2 with type of evasive AI malware is AI attack.

(2) When player 2 with type of evasive AI malware sticks to AI attack, we define E_{1n}(Non) (resp. E_{1n}(Non-AI)) as payoff of Non-defense (resp. Non-AI defense) strategy of player 1 with type of Non-AI detector. Also we define E_{1a}(Non) (resp. E_{1a}(AI)) as payoff of Non-defense (resp. AI defense) strategy of player 1 with type of AI detector. Clearly, E_{1n}(Non) > E_{1n}(Non-AI) holds since $-G_{aa} > -G_{aa}-C_{nd}$. If $I_{aa}=1$, E_{1a}(Non) > E_{1a}(AI) holds since $-G_{aa} > -G_{aa}-C_{ad}$. As a result, the optimal strategy of player 1 with type of Non-AI detector (resp. AI detector) is Non-defense (resp. Non-defense).

By (1), (2), the Theorem 1 is proved.

If both player 1 and player 2 stick to pure strategy Bayesian Nash equilibrium stated in Theorem 1, player 1 will do not defense against AI attack launched by player 2 under the condition of Theorem 1.

Theorem 2. A pure strategy Bayesian Nash equilibrium is ((Non-AI defense if Non-AI detector, AI defense if AI detector), Non-AI attack) under the condition that I_{na} =0 and ($\frac{G_{ad}+C_{na}}{G_{ad}-G_{nd}} < q < \frac{C_{aa}-C_{na}}{G_{aa}+G_{nd}}$ with I_{aa} =0 or $\frac{G_{aa}-C_{aa}+G_{ad}+C_{na}}{G_{ad}-G_{nd}} < q$ with I_{aa} =1) hold.

Proof. (1) We define E_2(Non-AI) and E_2(AI) as the expected payoff of Non-AI attack and AI-attack strategies of player 2 with type of evasive AI malware when player 1 with type of Non-AI detector sticks to Non-AI defense strategy and player 1 with type of AI detector sticks to AI defense strategy.

If I_{na} =0 and I_{aa} =0, $\frac{G_{ad}+C_{na}}{G_{ad}-G_{nd}} < q < \frac{C_{aa}-C_{na}}{G_{aa}+G_{nd}}$,

we have $E_2(\text{Non}-\text{AI}) = q(I_{na}G_{na} - (1-I_{na})G_{nd} - C_{na}) + (1-q)(-G_{ad} - C_{na}) = q(-G_{nd} + G_{ad}) - G_{ad} - C_{na}$, $E_2(\text{AI}) = q(G_{aa} - C_{aa}) + (1-q)(I_{aa}G_{aa} - (1-I_{aa})G_{ad} - C_{aa}) = q(G_{aa} + G_{ad}) - G_{ad} - C_{aa}$.

By the right above conditions of q, $E_2(\text{Non}-\text{AI}) > 0$ and $E_2(\text{Non}-\text{AI}) > E_2(\text{AI})$ holds.

If I_{na} =0 and I_{aa} =1, $\frac{G_{aa}-C_{aa}+G_{ad}+C_{na}}{G_{ad}-G_{nd}} < q$,

we have $E_2(\text{Non}-\text{AI}) = q(-G_{nd} + G_{ad}) - G_{ad} - C_{na}$, $E_2(\text{AI}) = G_{aa} - C_{aa}$. Since By the right above conditions of q, $E_2(\text{Non}-\text{AI}) > 0$, and $E_2(\text{Non}-\text{AI}) > E_2(\text{AI})$ hold. As a result, the optimal strategy of player 2 with type of evasive AI malware is Non-AI attack.

(2) When player 2 with type of evasive AI malware sticks to Non-AI attack, we define E_{1n}*(Non)* (resp. E_{1n}*(Non-AI)*) as payoff of Non-defense (resp. Non-AI defense) strategy of player 1 with type of Non-AI detector. Also we define E_{1a}*(Non)* (resp. E_{1a}*(AI)*) as payoff of Non-defense (resp. AI defense) strategy of player 1 with type of AI detector. Clearly, E_{1a}*(AI)* > E_{1a}*(Non)* holds since $G_{ad} > C_{ad}$. If I_{na} =0, E_{1n}*(Non-AI)* > E_{1n}*(Non)* holds since $G_{nd} > C_{nd}$. As a result, the optimal strategy of player 1 with type of Non-AI detector (resp. AI detector) is Non-AI defense (resp. AI defense).

By (1), (2), the Theorem 2 is proved.

If both player 1 and player 2 stick to pure strategy Bayesian Nash equilibrium stated in Theorem 2, player 1 with type of Non-AI detector will do Non-AI defense and player 1 with type of AI detector will do AI defense against Non-AI attack launched by player 2 under the condition of Theorem 2.

Theorem 3. A pure strategy Bayesian Nash equilibrium is ((Non-defense if Non-AI detector, AI-defense if AI detector), AI attack) under the condition that I_{aa} =0 and $\frac{G_{ad}+C_{aa}}{G_{aa}+G_{ad}} < q$ and $\frac{C_{aa}-C_{na}}{G_{aa}-G_{na}} < q$ hold.

Proof. (1) We define E_2(Non-AI) and E_2(AI) as the expected payoff of Non-AI attack and AI-attack strategies of player 2 with type of evasive AI malware when player 1 with type of Non-AI detector sticks to Non-defense strategy and player 1 with type of AI detector sticks to AI defense strategy.

If I_{aa} =0, $\frac{G_{ad}+C_{aa}}{G_{aa}+G_{ad}} < q$, $\frac{C_{aa}-C_{na}}{G_{aa}-G_{na}} < q$,

we have $E_2(AI) = q(G_{aa} - C_{aa}) + (1-q)(I_{aa}G_{aa} - (1-I_{aa})G_{ad} - C_{aa}) = q(G_{aa} + G_{ad}) -$

$G_{ad} - C_{aa}$

$$E_2(Non-AI) = q(G_{na} - C_{na}) + (1-q)(-G_{ad} - C_{na}) = q(G_{na} + G_{ad}) - G_{ad} - C_{na}$$

By the right above conditions of q, $E_2(AI) > 0$ and $E_2(AI) > E_2(Non-AI)$ holds. As a result, the optimal strategy of player 2 with type of evasive AI malware is AI attack.

(2) When player 2 with type of evasive AI malware sticks to AI attack, we define $E_{1n}(Non)$ (resp. $E_{1n}(Non\text{-}AI)$) as payoff of Non-defense (resp. Non-AI defense) strategy of player 1 with type of Non-AI detector. Also, we define $E_{1a}(Non)$ (resp. $E_{1a}(AI)$) as payoff of Non-defense (resp. AI defense) strategy of player 1 with type of AI detector. Clearly, $E_{1n}(Non) > E_{1n}(Non\text{-}AI)$ holds since $-G_{aa} > -G_{aa}-C_{nd}$. If $I_{aa}=0$, $E_{1a}(AI) > E_{1a}(Non)$ holds since $G_{ad}-C_{ad} > -G_{aa}$. As a result, the optimal strategy of player 1 with type of Non-AI detector (resp. AI detector) is Non-defense (resp. AI defense).

By (1), (2), the Theorem 3 is proved.

If both player 1 and player 2 stick to pure strategy Bayesian Nash equilibrium stated in Theorem 3, player 1 with type of Non-AI detector will do not defense and player 1 with type of AI detector will do AI defense against AI attack launched by player 2 under the condition of Theorem 3.

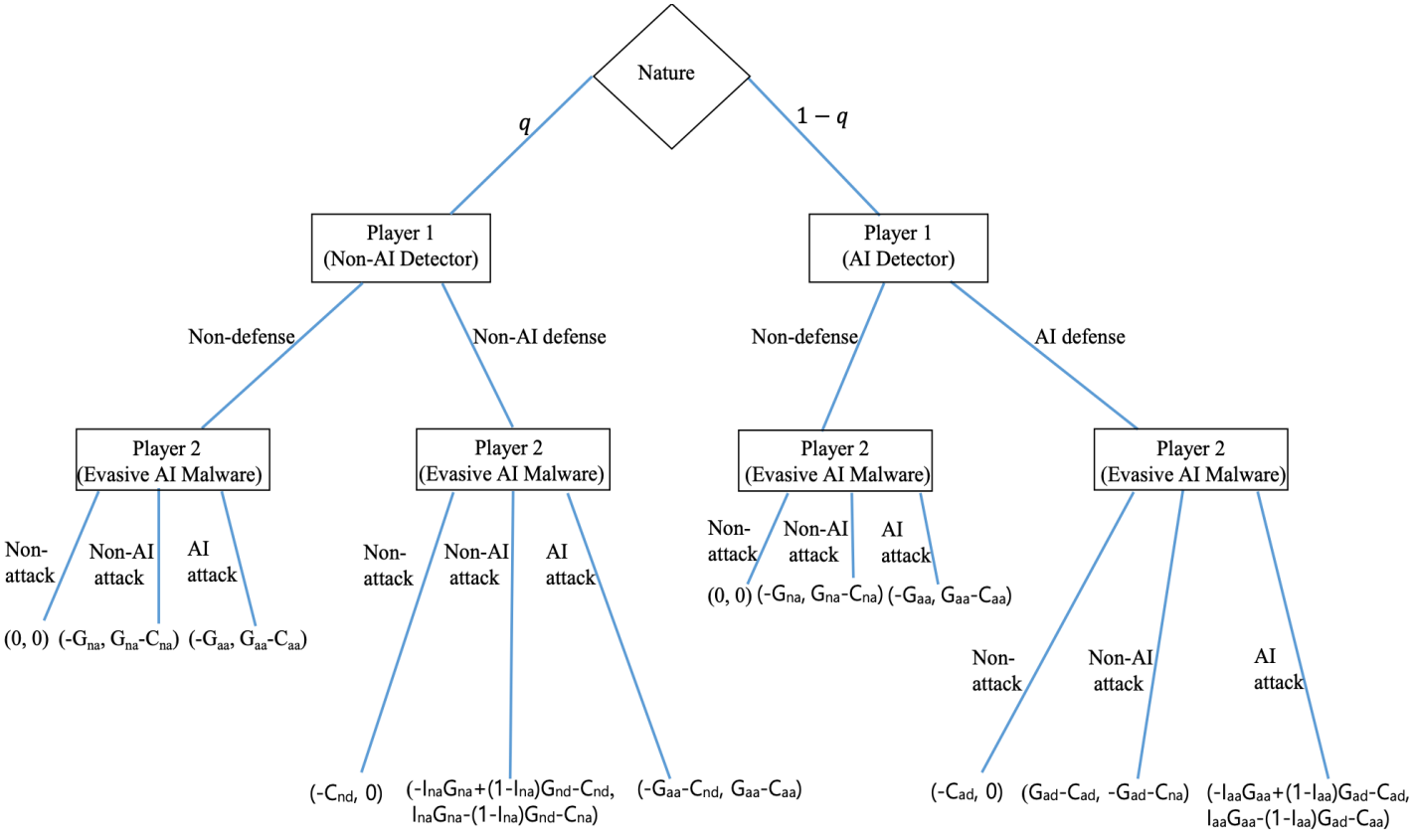

Figure 1: Extensive form of Bayesian Game in single stage.

4. Conclusion

In this work, we formulate a Bayesian signaling game for interaction analysis between evasive AI malware

and detector. We derive pure strategy Bayesian Nash Equilibria in a single stage game. As future work, we will plan to investigate whether mixed strategy Bayesian Nash Equilibria exist or not and explore the interaction analysis between evasive AI malware and detector in multi-stage game. In the sense that our devised work plays a role of game theoretic model for AI malware detection in IoT, it will have an impact upon how various AI malware detection schemes interact and defend against AI malware in IoT.

Acknowledgement

This work was supported by a research grant from Seoul Women's University (2024-0008).

References

[1] W. Diao, X. Liu, Z. Li, and K. Zhang. Evading Android Runtime Analysis Through Detecting Programmed Interactions. In *ACM WiSec*, 2016. DOI: https://doi.org/10.1145/2939918.2939926

[2] Jun-Won Ho . Game Theoretic Approach Toward Detection of Input-Driven Evasive Malware in the IoT. *TechRxiv.* September 29, 2022. DOI: https://doi.org/10.36227/techrxiv.19633677.v2

[3] D. Kirat, G. Vigna, C. Kruegel. BareCloud: Bare-metal Analysis-based Evasive Malware Detection. In *Usenix Security*, 2014.

[4] N. Miramirkhani, M. P. Appini, N. Nikiforakis, and M. Polychronakis. Spotless Sandboxes: Evading Malware Analysis Systems using Wear-and-Tear Artifacts. In *IEEE Symposium on Security and Privacy (SP)*, 2017. DOI: https://doi.org/10.1109/SP.2017.42

[5] J. Wampler, I. Martiny, and E. Wustrow. ExSpectre: Hiding Malware in Speculative Execution. In *Network and Distributed Systems Security* (NDSS) Symposium, 2019. DOI:https://doi.org/10.14722/ndss.2019.23409