On Adaptive Decision-Based Attacks and Defenses

Ilias Tsingenopoulos*, Vera Rimmer*, Davy Preuveneers*, Fabio Pierazzi†, Lorenzo Cavallaro‡, Wouter Joosen*
*KU Leuven, †King's College London, ‡University College London

***Abstract*—Despite considerable efforts on making them robust, real-world AI-based systems remain vulnerable to decision-based attacks, as definitive proofs of their operational robustness have so far proven intractable. The canonical approach in robustness evaluation calls for adaptive attacks, that is with complete knowledge of the defense and tailored to bypass it. We introduce a more expansive notion of adaptive and show how not only attacks but also defenses should optimize through the competitive game they form. To reliably measure robustness, it is important to evaluate against realistic and worst-case attacks. We augment attacks themselves *and* the evasive arsenal at their disposal through adaptive control, while the same can be done for defenses. We argue that active defenses, which control how the system responds, are a necessary complement to model hardening when facing decision-based attacks; then show how these defenses can be circumvented by adaptive attacks, only to finally elicit active *and* adaptive defenses. We assert that AI-enabled adversaries pose a considerable threat to black-box ML systems, rekindling the proverbial arms race where defenses *have* to be AI-enabled too.**

1. Introduction

AI models are predominantly trained, validated, and deployed with little regard to their correct functioning under adversarial activity, often leaving safety, ethical, and broader societal impact considerations as an afterthought. Adversarial contexts further aggravate the typical generalization challenges that these models face with threats beyond model evasion (extraction, inversion, poisoning [19]) while the systems they enable often expose interfaces that can be queried and used as adversarial "instructors". Scoping on model evasion, the most reliable mitigation to date is adversarial training [25], [36], an approach not without limitations as these models often remain irreducibly vulnerable at deployment, particularly against black-box, decision-based attacks [6], [10], [38].

In adversarial machine learning (AML) evaluating the robustness of defenses against oblivious, non-adaptive, and thus suboptimal attackers is inherently problematic [14], [35]. Here we expand the conventional notion of adaptive, from *adapted* attacks that have an empirical configuration to bypass the defense, to include the capability to *self-adapt*, where attacks adaptively control their parameters *and* evasive actions together in response to how the model under attack and its defenses respond [2]. We propose that self-adaptive adversaries can modify their own policies through reinforcement learning (RL) to become both optimal *and* evade active detection. Notably, this can be performed in a gradient-based manner even in fully black-box contexts [1], a capability that *properly reflects* the level of adversarial threat and does not overestimate the empirical robustness; attackers will compute gradients after all.

All such attacks exhibit a behavior at-the-interface that can be described as adversarial itself, a generalization that subsumes adversarial examples and opens a path towards novel defenses and mitigations. Aside from making the underlying models more robust, this behavior can be countered as such rather than relying on hardened models exclusively. As models cannot update their decision boundary in an online manner and in response to adversarial activity on their interface, there *has* to be a complement to model hardening: for instance *active* defenses such as rejection or misdirection [4], [11], [31]. We contend that robust evaluations cannot consider attacks or defenses in isolation, and that each should have the capability to modify their operation through interaction and in direct response to other agency in the environment.

2. On Being Adaptive

While adversarial attacks have been extensively researched in both white and black-box manner, defenses have predominantly focused on white-box contexts [25], [36]. As the black-box setting discloses considerably less information, a seemingly intuitive conclusion is that white-box defenses should suffice for the black-box case too. Yet black-box attacks like [6], [10] have been highly effective against a wide range of defenses like *gradient masking* [3], *preprocessing* [9], [29], and *adversarial training* [25]. The vast majority of adversarial defenses provide either limited robustness or are eventually evaded by adapted attacks [35]. Characteristically, preprocessing defenses are often bypassed by expending queries for reconnaissance [32]. The partial exception to this rule is adversarial training [25], briefly described in Appendix A.

The correct way to evaluate a defense is against *adaptive* attacks: attacks with explicit knowledge of the inner workings of a defense [35]. If however model hardening is the defensive counterpart to white-box attacks, an active defense like stateful detection is the counterpart to decision-based attacks, and also the *necessary* complement to model hardening. All decision-based attacks share an inherent sequentiality that can be useful in devising defenses against them, like Blacklight does by rejecting queries through

quantization and hashing [22]. This defense was recently bypassed by adapting existing attacks through rejection sampling [15]. A simple adaptation however would put the efficacy of the latter in question: what if Blacklight did not disclose *more* information through the rejection signal? An adversary could *still* adapt and devise a way to discern when rejection takes place. The above are canonical examples of what is meant as adaptive in AML.

At the same time, the level of threat attacks pose is often unclear or not thoroughly evaluated. As demonstrated by Croce et al. [14], it is very common that the parameters of an attack are suboptimal, leading to *underestimating* their performance and thus *overestimating* the claimed degree of robustness. This can be further aggravated in black-box contexts, where the attacker is largely oblivious of any pre-processing or active defenses. The effectiveness of attacks rests on the ability to adapt the policies that govern both their operation and their evasive capabilities *in tandem*. Here we expand on the notion of adaptive, as it is conventionally understood, to include *adaptive control*: the ability of a system to **self-adapt** and reconfigure itself in response to changes in the dynamics of the environment in order to achieve optimal behavior [2].

Typically, what is to be controlled is known in advance and well-defined. The moment however we consider adaptive evaluations, *new* controls are immediately implied, like rejection sampling in Blacklight. To flesh out the twofold meaning of adaptive, one has to *both* imagine new knobs [21], *and* discover their correct configuration. The invention of knobs, a faculty strictly human so far, is a way to impart controllability to the task: in our case the instruments to bypass an existing defense. We conceptualize this expanded definition of adaptive, essential for having accurate evaluations in AML research, in Figure 1.

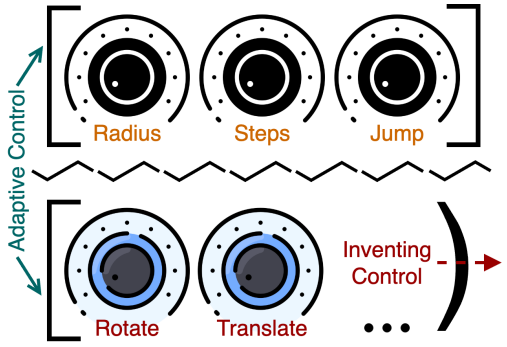

Figure 1. In AML adaptive means to invent new knobs that can bypass a defense (lower open set); in control theory it means the precise tuning of all known knobs. In this work, we reformulate adaptive to signify **both**. In HSJA [10] for example, radius, steps, and jumps are parameters of the attack, while rotate and translate are evasive transformations.

3. Adversarial Markov Games

The most compelling threat for deployed ML systems are hard-label, decision-based attacks like Boundary Attack [6], HSJA [10], Guessing Smart (BAGS) [8], Sign-Opt [12], Policy-driven (PDA) [38], QEBA [23], and SurFree [26], which are becoming increasingly effective. In HSJA for example, its optimization is guaranteed to converge to a stationary point, which given typical values on perturbation imperceptibility translates to near-perfect attack success rates, even against **adversarially trained** models. The limitations of adversarial training against decision-based attacks can be attributed to the fundamentally out-of-distribution (OOD) nature of adversarial examples, as that makes the saddle point optimization of Eq. (1) intractable to solve exhaustively. Additionally, it is challenging to incorporate decision-based attacks *during* stochastic gradient descent: as approaches that navigate the decision boundary, the further the latter is from convergence, the less effective the attack is.

Decision-based attacks search for the **optimal parameters** of the adversarial policy, those that minimize the perturbation in expectation; given its dimensionality however, it can be intractable to learn a policy that modifies the input space directly [28]. Instead, given a well-defined set of controls we can formulate the adversarial task as a Markov Decision Process (MDP) to be solved and learn an optimal policy that minimizes the perturbation *and* evades detection.

In AI-enabled systems, the best practice is to freeze the model after validation so that no novel issues are introduced by retraining. While this is representative of real-world settings, it is also what enables adversaries to discover adversarial examples that could not be accounted for in advance. The existence of a adversarial policy introduces however a behavior which can be observed and utilized by a defensive methodology. Consequently, model-hardening approaches like adversarial training are *necessary* but also *insufficient* in defending against decision-based attacks.

Claim 1. *Given an adversarially trained model $\mathcal{M}$, to fully defend against decision-based attacks, two additional capabilities are necessary: a) a decision function other than the most probable class, and b) additional stateful context upon which this decision is taken.*

Intuitively, if the model *always* responds truthfully, the adversary will be able to accurately execute its policy and converge to the optimal adversarial; secondly, the model should be able to differentiate between two, otherwise identical, queries when one is part of an attack and when is not. Classification with rejection or intentional misdirection can be such decision functions. The former has manifested in the form of conformal prediction or learning with rejection [4], [13]; while misdirection has emerged as a technique in adversarial RL and cybersecurity domains [16], [31]. While adversarially training can partially resist decision-based attacks, the manner in which the model responds has complementary potential. This gap between the empirical and theoretically possible robustness to decision-based attacks is the locus where a, distinct from model hardening, active defense can emerge.

Active defenses have immediate implications on the attacks themselves however. Deployed defenses are by definition fixed, which makes the environment dynamics sta-

tionary. Bypassing the defense can then *also* be formulated as an MDP to be solved. In such a two-player, zero-sum game, following a stationary policy becomes *exploitable* through the reward obtained by the opponent [34]. Active defenses, a consequence of decision-based attacks, entail adaptive adversaries.

Claim 2. *Against an active defense $\pi_\phi^{\mathcal{D}}$, a decision-based attack following a non-adaptive and stationary adversarial policy will perform arbitrarily worse in expectation.*

Consider now an active similarity-based defense. In the twofold meaning of adaptive we introduced, inventing control implies the *potential* to bypass; adaptive control implies strategy instead, the online configuration of the available tools for evasion [1]. Notably, this optimization can be *fully* gradient-based despite the discrete and black-box nature of the adversarial task [33]. Adaptively controlling attacks with RL can recover the *gradient-based* solvability of the black-box optimization task despite *neither* the active defense *nor* the model itself being accessible in closed-form.

Postulate 1 (Adversarial Policy Gradient)**.** *Given model $\mathcal{M}$ with an active defense $\pi_\phi^{\mathcal{D}}$, adversary policy $\pi_\theta^{\mathcal{A}}$ that generates queries x_t, and a reward $\mathcal{R}(\tau) \in \{0, 1\}$ reflecting failure or success respectively in episode τ, the optimal attack policy is obtained via the gradient of the expected reward $\mathbb{E}_{\pi_\theta^{\mathcal{A}}}[\mathcal{R}(\tau)]$.*

An intuitive understanding of this postulate can be derived from the Policy Gradient Theorem [33], and its formulation is general to cover *any* defense mechanism upon which examples are rejected, for example by using explanations [27]. We note here that evasive transformations, which the model (but not the defense) is invariant to, interfere with the perturbations from the adversarial policy itself: the performance and evasiveness of an attack are typically in a natural trade-off. These transformations can be considered as set of additional controls, and like attack parameters they themselves can be underperforming out-of-the-box [14]. Thus the combined control of attack and evasion parameters is a *prerequisite* to properly assess the strength of a defense: their empirical configuration is often suboptimal. This trade-off illustrates why the twofold definition of adaptive is necessary in AML evaluations: first to impart controllability to the task through the definition of *what* can be controlled, and then to find the optimal execution of the attack. The last piece of the puzzle is turning active defenses also adaptive.

Claim 3. *An active defense $\pi_\phi^{\mathcal{D}}$ achieves its optimum, i.e. maximizes the expectation on perturbation, by adapting its policy against a stationary attacker policy.*

As offensive and defensive policies are strictly competitive, we can define the reward P of the defensive policy as $P(\tau) = \neg\mathcal{R}(\tau)$, then by making $\pi_\theta^{\mathcal{A}}$ stationary and $\pi_\phi^{\mathcal{D}}$ adaptive in Postulate 1, we can reason that the optimal defensive policy is determined also via the gradient of its expected reward. When offensive or defensive methodologies become adaptive, their environments become in turn non-stationary [20], putting further pressure on the IID foundations ML builds on. This multi-agent interaction constitutes a competitive and sequential zero-sum game [5], [18], [24] that we describe as an Adversarial Markov Game (AMG). A more formal treatment is included in Appendix B.

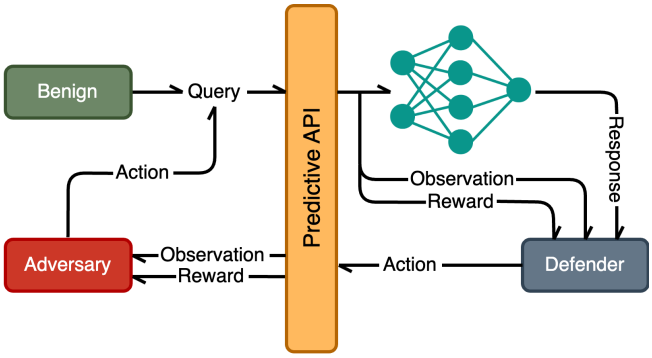

Figure 2. Schematic model of an AMG environment. Due to the inherent uncertainty of behavior at either side of the interface, it is a partially observable environment mirrored for each agent where one's decisions become the other's observations.

4. Discussion

With AI-enabled decision-making becoming pervasive in domains like governance, finance, employment, and of course cybersecurity, more and more decisions are delegated to AI which becomes increasingly accountable for upholding safety and ethical constraints. As trustworthy AI is vital for the healthy functioning of whole ecosystems, we highlight security risks and potential mitigations in these inherently black-box environments. While adversarial training remains the most reliable defense, the amount of robustness it imparts will vary and even be insufficient as AI-enabled systems are susceptible to adaptive adversaries that *devise* new evasive techniques and control them *jointly* with other attack parameters. This can been achieved in the *fully* black-box case and against *active* defenses; our Adversarial Policy Gradient indicates that any combination of adversarial goals – be it performance, stealthiness, disruption – can be optimized in a *gradient-based* manner and it is straightforward to generalize to any domain or modality.

In self-adaptive, we introduce a novel twofold definition of adaptive: both devising new methods of outmaneuvering opponents *and* adapting one's operating policy with respect to other agency in the environment. The AMG formulation we introduce helps us reason on and assess the vulnerabilities of AI-based systems by disentangling the inherently complex and non-stationary task of learning in the presence of competing agency; by modeling it as a fixed part of the environment, we can simplify the task by computing a best response against the observed behavior. This is an important outcome for cybersecurity domains: as long as proper threat analysis is carried out, one can readily employ RL algorithms in order to devise optimal defenses; but only after they devised optimal attacks too.

Acknowledgment

This research is partially supported by the Research Fund KU Leuven, the Flemish Research Programme Cybersecurity, the European Commission through the Horizon Europe project KINAITICS[1] under grant agreement 101070176, and by the UK EPSRC Grant EP/X015971/1.

References

[1] S. V. Albrecht and P. Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

[2] K. J. Åström and B. Wittenmark. *Adaptive control*. Courier Corporation, 2013.

[3] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283, 2018.

[4] F. Barbero, F. Pendlebury, F. Pierazzi, and L. Cavallaro. Transcending transcend: Revisiting malware classification in the presence of concept drift. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 805–823. IEEE, 2022.

[5] J. Bose, G. Gidel, H. Berard, A. Cianflone, P. Vincent, S. Lacoste-Julien, and W. Hamilton. Adversarial example games. *Advances in neural information processing systems*, 33:8921–8934, 2020.

[6] W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018.

[7] M. Brückner and T. Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD conference on Knowledge discovery and data mining*, pages 547–555, 2011.

[8] T. Brunner, F. Diehl, M. T. Le, and A. Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4958–4966, 2019.

[9] J. Byun, H. Go, and C. Kim. On the effectiveness of small input noise for defending against query-based black-box attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3051–3060, 2022.

[10] J. Chen, M. I. Jordan, and M. J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020.

[11] S. Chen, N. Carlini, and D. Wagner. Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, pages 30–39, 2020.

[12] M. Cheng, S. Singh, P. H. Chen, P.-Y. Chen, S. Liu, and C.-J. Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2019.

[13] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016.

[14] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[15] R. Feng, A. Hooda, N. Mangaokar, K. Fawaz, S. Jha, and A. Prakash. Stateful defenses for machine learning models are not yet secure against black-box attacks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 786–800, 2023.

1. https://kinaitics.eu

[16] A. Gleave, M. Dennis, N. Kant, C. Wild, S. Levine, and S. Russsell. Adversarial policies: Attacking deep reinforcement learning. *In Proc. ICLR-20*, 2020.

[17] A. Greenwald, J. Li, and E. Sodomka. Solving for best responses and equilibria in extensive-form games with reinforcement learning methods. In *Rohit Parikh on Logic, Language and Society*, pages 185–226. Springer, 2017.

[18] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[19] Y. He, G. Meng, K. Chen, X. Hu, and J. He. Towards security threats of deep learning systems: A survey. *IEEE Transactions on Software Engineering*, 48(5):1743–1770, 2020.

[20] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. de Cote. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2017.

[21] D. R. Hofstadter. *Metamagical themas: Questing for the essence of mind and pattern*. Hachette UK, 2008.

[22] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng, and B. Y. Zhao. Blacklight: Scalable defense for neural networks against {Query-Based}{Black-Box} attacks. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2117–2134, 2022.

[23] H. Li, X. Xu, X. Zhang, S. Yang, and B. Li. Qeba: Query-efficient boundary-based blackbox attack. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[24] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings*. Elsevier, 1994.

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[26] T. Maho, T. Furon, and E. Le Merrer. Surfree: a fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10430–10439, 2021.

[27] M. Noppel and C. Wressnegger. Sok: Explainable machine learning in adversarial environments. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 21–21. IEEE Computer Society, 2023.

[28] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro. Intriguing properties of adversarial ml attacks in the problem space. In *2020 IEEE symposium on security and privacy (SP)*, pages 1332–1349. IEEE, 2020.

[29] Z. Qin, Y. Fan, H. Zha, and B. Wu. Random noise defense against query-based black-box attacks. *Advances in Neural Information Processing Systems*, 34:7650–7663, 2021.

[30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

[31] S. Sengupta and S. Kambhampati. Multi-agent reinforcement learning in bayesian stackelberg markov games for adaptive moving target defense. *arXiv e-prints*, pages arXiv–2007, 2020.

[32] C. Sitawarin, F. Tramèr, and N. Carlini. Preprocessors matter! realistic decision-based attacks on machine learning systems. *arXiv preprint arXiv:2210.03297*, 2022.

[33] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer, 1999.

[34] F. Timbers, N. Bard, E. Lockhart, M. Lanctot, M. Schmid, N. Burch, J. Schrittwieser, T. Hubert, and M. Bowling. Approximate exploitability: Learning a best response. *IJCAI, Jul*, 2022.

[35] F. Tramer, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33:1633–1645, 2020.

[36] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu. On the convergence and robustness of adversarial training. In *International Conference on Machine Learning*, pages 6586–6595. PMLR, 2019.

[37] Y. Wen, Y. Yang, R. Luo, J. Wang, and W. Pan. Probabilistic recursive reasoning for multi-agent reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[38] Z. Yan, Y. Guo, J. Liang, and C. Zhang. Policy-driven attack: learning to query for hard-label black-box adversarial examples. In *International Conference on Learning Representations*, 2020.

Appendix A.

Given dataset $D = (x_i, y_i)_{i=1}^n$ with classes C where $x_i \in \mathbb{R}^d$ is a clean example and $y_i \in 1, ..., C$ is the associated label, the objective of adversarial training is to solve the following *min-max* optimization problem:

$$\min_{\phi} \mathbb{E}_{i \sim D} \max_{\|\delta_i\|_{L_p} \leq \epsilon} \mathcal{L}(h_\phi(x_i + \delta_i), y_i) \tag{1}$$

where $x_i + \delta_i$ is an adversarial example of x_i, $h_\phi : \mathbb{R} \to \mathbb{R}^{\mathbb{C}}$ is a hypothesis function and $\mathcal{L}(h_\phi(x_i + \delta_i), y_i)$ is the loss function for the adversarial example $x_i + \delta_i$. The inner maximization loop finds an adversarial example of x_i with label y_i for a given L_p-norm (with $L_p \in \{0, 1, 2, \inf\}$), such that $\|\delta_i\|_l \leq \epsilon$ and $h_\phi(x_i + \delta_i) \neq y_i$. The outer loop is the standard minimization task typically solved with stochastic gradient descent. While the convergence and robustness properties of adversarial training have been investigated through the computation of the inner maximization step and by interleaving normal and adversarial training [36], the min-max principle is conspicuous: minimize the possible loss for a worst case (max) scenario.

Appendix B.

As adaptive decision-based attacks and defenses are logical consequences of each other, by composing them we can form a turn-taking competitive game. A precise game-theoretic formulation however requires the exact analytical description of the whole environment: the model, the players and their utility functions, as well as the permitted interactions and the transition dynamics, something exceedingly intractable in most cybersecurity environments. Model-free methods however can learn optimal offensive and defensive responses directly through interaction with the environment [30], [31], obviating the need to learn a model of it or to find *exact* solutions to the bilevel optimization task of adversarial training that is NP-hard to solve [7].

We model AMGs after Turn-Taking Partially-Observable Markov Games (TT-POMGs), introduced by Greenwald et al. [17]. TT-POMGs are a generalization of Extensive-Form Games (EFGs), widely used representations for non-cooperative, sequential decision-making games of imperfect or incomplete information. A useful property of TT-POMGs is that they can be transformed to equivalent belief state MDPs, significantly simplifying their solution. By folding other agents strategies into the transition probabilities and the initial probability distribution of the game, an optimal policy computed in the resulting MDP will correspond to the best-response strategy in the original TT-POMG. The congruence between TT-POMGs and MDPs is useful also for its practical implications in the security of ML-based systems: provided that adversarial agents and their capabilities can be identified through rigorous threat analysis, computing the best response strategy in the simulated environment will correspond to the **optimal** defense.

The goal of each player in an AMG – depicted in Figure 2 – is to determine a policy that maximizes their expected reward. When a player employs a stationary policy, the AMG reduces to a belief-state MDP where others interact with a fixed environment. Formally, we represent AMG as a tuple $\langle i, S, O, A, \tau, r, \gamma \rangle$

- $i = \{\mathcal{D}, \mathcal{A}\}$ are the players, where $\mathcal{D}$ denotes the defender and $\mathcal{A}$ denotes the adversary. In our model, benign queries are modeled as moves by nature.
- S is the full state space of the game, while $O = \{O^{\mathcal{D}}, O^{\mathcal{A}}\}$ are partial observations of the full state for each player.
- $A = \{A^{\mathcal{D}}, A^{\mathcal{A}}\}$ denotes the action set of each player.
- $\tau(s, a^i, s')$ represents the transition probability to state $s' \in S$ after player i chooses action a^i.
- $r = \{r^{\mathcal{D}}, r^{\mathcal{A}}\} : O^i \times A^i \to \mathbb{R}$ is the reward function where $r^i(s, a^i)$ is the reward of player i if in state s action a^i is chosen.
- $\gamma^i \in [0, 1)$ is the discount factor for player i.

The game is sequential and turn-taking, so each player i chooses an action a from A^i which subsequently influences the observations of others. As without implausible assumptions one cannot assume access to the exact state of other agents, each state is a partial observation of the complete state of the full game. When a player employs a fixed policy, the AMG reduces to a belief-state MDP where the other interacts with a stationary environment. Considering such policies as part of the environment is equivalent to *0th* level recursive reasoning in the study of opponent modeling: the agent models how the opponent behaves based on the observed history, but *not* how the opponent *would* behave based on how the agent behaves [1], [37].

AMGs can be solved with single-agent RL; as AI-enabled systems proliferate we expect that more involved recursive reasoning and explicit opponent modeling will prove essential. The most compelling and formidable challenge however remains the automation of adaptive evaluations in AML, by inventing instruments of bypassing defenses and imparting controllability to the adversarial task.