Adversarial Markov Games: On Adaptive Decision-Based Attacks and Defenses

Anonymous Author(s)

***Abstract*—Despite considerable efforts on making them robust, real-world ML-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, with complete knowledge of the defense and tailored to bypass it. In this study, we introduce a more expansive notion of being adaptive and show how attacks but also defenses can benefit by it *and* by learning from each other through interaction.**

We propose and evaluate a framework for adaptively optimizing black-box attacks and defenses against each other through the competitive game they form. To reliably measure robustness, it is important to evaluate against realistic and worst-case attacks. We thus augment both attacks *and* the evasive arsenal at their disposal through adaptive control, and observe that the same can be done for defenses, before we evaluate them first apart and then jointly under a multi-agent perspective. We demonstrate that active defenses, which control how the system responds, are a necessary complement to model hardening when facing decision-based attacks; then how these defenses can be circumvented by adaptive attacks, only to finally elicit active *and* adaptive defenses.

We validate our observations through a wide theoretical and empirical investigation to confirm that AI-enabled adversaries pose a considerable threat to black-box ML-based systems, rekindling the proverbial arms race where defenses *have* to be AI-enabled too. Our approach outperforms the state-of-the-art adaptive attacks and defenses, while bringing them together to render effective insights over the robustness of real-world deployed ML-based systems.

I. 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 [26]) while the systems they enable often expose interfaces that can be queried and used as adversarial "instructors", like in constructing adversarial malware against existing ML-based malware detection [3], [19]. Scoping on model evasion, the most reliable mitigation to date is adversarial training [33], [50], an approach not without limitations as these models often remain irreducibly vulnerable at deployment, particularly against black-box, decision-based attacks [8], [14], [52]. Nevertheless, 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.

Adversarial behavior is a temporal extension of adversarial examples, perhaps not malicious or harmful in isolation, yet part of an attack as it unfolds over time; it is also the canonical description of adversarial examples in domains like dynamic malware analysis and adversarial RL [48], [22]. 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 [6], [42], [15].

In this study we address a crucial gap, as in adversarial machine learning (AML) evaluating the robustness of defenses against oblivious, non-adaptive, and therefore suboptimal attackers is inherently problematic [47], [18]. 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 [4]. We demonstrate theoretically and empirically how self-adaptive attacks can modify their policies through reinforcement learning (RL) to become both optimal *and* evade active detection. Notably, we find that this can be performed in a gradient-based manner even in fully black-box contexts [III.1], a capability that *properly reflects* the level of adversarial threat and does not overestimate the empirical robustness; attackers might compute gradients after all.

This capability, however, enables active defenses in turn, as through rigorous threat modeling and by simulating self-adaptive attackers their full potential is uncovered and effective counter-policies can be learned. To express the call for adaptive evaluations in AML differently, a defense can only be as good as its adversary. This mutual interdependence generates the necessity for *both* attacks and defenses being self-adaptive, as well as the competitive, zero-sum game they jointly form. In summary, our research engages from two related perspectives: a) the optimality of decision-based attacks, and b) defenses against them, resulting in the following contributions:

- We demonstrate that active defenses against decision-based attacks are a *necessary* but *insufficient* complement to model hardening. Active defenses are inevitably bypassed by self-adaptive attackers however, and necessitate **self-adaptive** defenses too.
- To facilitate reasoning on adaptive attacks and defenses, we introduce a unified framework called "Adversarial Markov Games" (**AMG**). We demonstrate how adversaries can optimize their policy and evade active detection *at the same time*; as a counter, we propose a novel active

defense and employ RL agents to **adapt** and optimize both. For reproducibility and follow-up work, we open-source our code[1].

- In an extensive empirical evaluation on image classification and across various adversarial settings, we validate our theoretical analysis and show that self-adaptation through RL **outperforms** the baseline attacks, model hardening defenses like adversarial training, and notably **both** the state-of-the-art adaptive attacks and stateful defenses.

Our work indicates that in the domain of black-box, decision-based AML, robust evaluations *should* go a step further than including adapted attacks: both attacks and defenses should have the capability to modify their operation through interaction and in direct response to other agency in the environment. The remainder of the paper is structured as follows: Section II provides the necessary background on the domain and reviews the related work. Section III introduces and motivates our theoretical analysis of robustness under decision-based attacks. Section IV elucidates the threat model and the concrete design choices. In Section V we elaborate on our experimentation and analyze our results. We conclude with Section VI where we discuss insights, limitations and challenges.

II. Preliminaries

In this work, we focus on the category of adversarial attacks known as **decision-based**, a subset of query-based attacks that operate solely on the **hard-label** outputs of the model and are regarded as a highly realistic and pervasive threat in AI-based cybersecurity environments. Despite the lack of the closed-form expression of the model under attack, given enough queries the effectiveness of such black-box attacks can match and even surpass that of white-box techniques like C&W [13].

A. Attacks & Mitigations

While adversarial attacks have been extensively researched in both white and black-box contexts, defenses have predominantly focused on the white-box context [33], [50]. 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 [8], [14] have shown to be highly effective against a wide range of defenses like *gradient masking* [5], *preprocessing* [39], [12], and *adversarial training* [33]. The vast majority of adversarial defenses provide either limited robustness or are eventually evaded by adapted attacks [47]. Characteristically, preprocessing defenses are often bypassed by expending queries for reconnaissance [43].

The partial exception to this rule is adversarial training [33]. 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 [50], the min-max principle is conspicuous: minimize the possible loss for a worst case (max) scenario.

B. Stateful Defenses

All decision-based attacks share properties that can be useful in devising defenses against them, *on top* of adversarial training. Such a property is their inherent sequentiality: by following an attack policy towards the optimal adversarial example, the generated candidates are correlated. This might not be the case for the queries themselves however, as the adversary might employ transformations the model is invariant to, like the query blinding strategy in [15]. This work is also the first to employ a *stateful* defense against query-based attacks. Another one is PRADA [29], a defense built against model extraction but which also works against model evasion. The efficacy of these approaches however rests on the assumption that queries can be consistently linked (through metadata like IP or account, cf. Table I) to uniquely identifiable actors – that also show limited to no collaboration – so that a buffer of queries can be built for each.

This limitation was recently addressed, together with the ensuing scalability issues in the Blacklight defense, by resourcefully employing hashing and quantization [30]. It remains a similarity-based defense however, thus vulnerable to circumvention if an adversary can find a query generation policy that preserves attack functionality while evading detection; the recent OARS work achieved this by adapting existing attacks through the rejection signal Blacklight returns [20]. Ultimately, any (stateful) defense has to balance the trade-off between robust and clean accuracy; as we demonstrate in this work, this trade-off can be representative only if the adversary has exhausted their adaptive offensive capabilities.

C. On Being Adaptive

The correct way to evaluate any proposed defense is against *adaptive* attacks, that is with explicit knowledge of the inner mechanisms of a defense [47]. In computer security this is known as the stipulation that security through obscurity does not work, as the robustness of defenses should not rely on keeping their way of functioning secret. If model hardening – for instance by adversarial training – is the defensive counterpart to white-box attacks, active defenses like stateful

[1]https://anonymous.4open.science/r/AMG-AD16

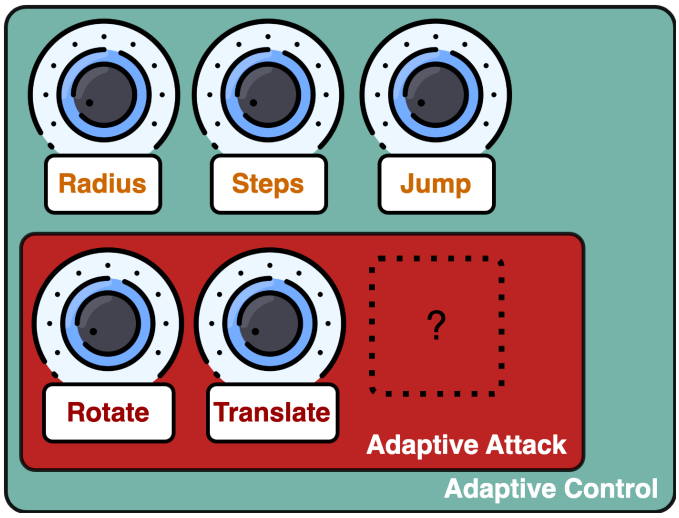

Fig. 1. In AML, adaptive attacks are those with the capabilities (knobs) to bypass a defense (red set); adaptive control is rather the precise tuning of all the known knobs. For decision-based environments, we can reformulate adaptive so that it signifies **both**. For instance in HSJA [14], radius, steps, and jump are parameters of the attack, while rotate and translate are transformations that can evade a similarity-based defense.

detection are the counterpart to decision-based attacks, and as we will further demonstrate, also the *necessary* complement to hardening a model against them.

At the same time, the level of threat that attacks pose is often unclear or not thoroughly evaluated. Previous work has demonstrated that loss functions and parameters of attacks are often suboptimal, leading to *underestimating* their performance and thus *overestimating* the claimed degree of robustness [18], [38]. This underestimation is further aggravated in decision-based contexts, where the attacker is largely oblivious of any preprocessing or active defenses the black-box system might have. The practical effectiveness of attacks therefore rests on the ability to adapt the policies that govern their operation and their evasive capabilities *in tandem*.

In AML, "adaptive" by convention refers to attacks with full knowledge of how a defense works and the tools to bypass it; we denote such attacks as **adapted**. In our work, we expand the term to include *adaptive control*, defined as the ability of a system to **self-adapt**: *automatically* reconfigure itself in response to changes in the dynamics of the environment in order to achieve optimal behavior [4]. One can think of adaptive control in the sense "attack optimization" is used by Pintor et al. [38], but for black-box systems. Typically, what is to be controlled is well-defined and known in advance. The moment however we consider adaptive evaluations, *new* controls are potentially implied: in a similarity-based defense for instance, such controls would be input transformations the model is invariant to. To flesh out the twofold meaning of adaptive, one has to *both* invent new knobs [28] (conventional understanding of adaptive, and still very hard to automate), *and* dynamically control their correct configuration that would lead to the optimal result (self-adaptive). We conceptualize this more general definition of adaptive, essential for having accurate evaluations against decision-based attacks, in Figure 1.

D. Research Gap

Prior work has focused on *adapted* attacks, which incorporate general knowledge of any defenses and empirically configured to evade it [13], [14], [8]. Defenses also follow the same adapted paradigm of empirically defined and fixed parameters [15], [30]. Our observation is that neither of them are formalized or performed in a fully adaptive manner, that is in response to how they influence their environment and with respect to other adaptive agents in it, with clear limitations when the latter is a given, e.g. in cybersecurity. To bridge this gap, we provide a theoretical treatment and empirical study of existing and novel methodologies adapting through direct interaction with their environment, denoting them as **self-adaptive**.

Our work builds on a long line of prior research that focuses on both sides of the competition between adversaries and defenses. Carlini and Wagner [13] show that evaluating existing attacks out-of-the-box is insufficient and that adapted white-box attackers can break defensive distillation. Bose et al. [7] propose Adversarial Examples Games (AEG), a zero-sum game between a white-box attacker and a local surrogate of the target model family. At the equilibrium the attacker can generate adversarial examples that have a high success rate against models from the same family, constituting a zero-query, non-interactive approach for generating transferable adversarial examples. Pal et al. [35] propose a game-theoretic framework for studying white-box attacks and defenses that occur in equilibrium. Feng et al. [20] introduce OARS: adaptive versions of existing attacks that bypass Blacklight [30], the state-of-the-art stateful defense. To function, OARS presupposes the rejection signal that a defense like Blacklight returns; a strong assumption that as we show in this work does not have to hold for stateful defenses. As we demonstrate in section V and Table III, Blacklight can be bypassed without assuming rejection, while the novel stateful defense we introduce can fully withstand the OARS adaptive attack.

As the most relevant and representative threat against real-world AI systems, in this work we scope on decision-based, interactive attacks and defenses. We contribute a theoretical and practical framework for self-adaptation, under which the full extent of the offensive and thus also the defensive potential is properly assessed. In the remainder of the paper the term **"adaptive"** subsumes adaptive control, and use it interchangeably with **"self-adaptive"**. For what is conventionally known as adaptive evaluations in AML, we use the term **"adapted"**. For ease of comparison, in Table I we highlight the most important aspects of our work as the synthesis of adaptive black-box attacks and defenses in a unified framework, and situate it with respect to other prominent and SOTA works in AML. Note the importance for an attack to function without assuming rejection, and respectively for a defense to function without access to query metadata like UIDs.

TABLE I

PROMINENT DECISION-BASED ATTACKS AND DEFENSES AND THEIR INDIVIDUAL ASPECTS. WHILE ALL WORKS ARE SITUATED EITHER ON THE OFFENSIVE OR THE DEFENSIVE SIDE, OURS BRINGS THESE TWO TOGETHER.

Work	Offensive				Defensive			
	Optimized	Evasive	Adaptive	¬Rejection	Active	Adaptive	¬Metadata	Misdirection
Boundary (2018) [8]	○	○	○	●	○	○	—	—
BAGS (2018) [11]	○	○	○	●	○	○	—	—
HSJA (2020) [14]	●	○	○	●	○	○	—	—
OARS (2023) [20]	●	●	◐	○	●	○	—	—
Adv. Training (2017) [33]	●	○	○	—	○	○	—	—
Stateful (2020) [15]	●	●	○	○	●	○	○	○
Blacklight (2022) [30]	●	●	○	○	●	○	●	○
AMG – Our work	●	●	●	●	●	●	●	●

III. THEORETICAL FRAMEWORK

In this work we engage from two perspectives that are interrelated: a) thwarting decision-based attacks, and b) adapting attacks and evasive capabilities *in tandem*. Evaluating non-adaptive, especially in the expansive sense we outlined, attacks or defenses renders results unreliable and incomplete [47]. When offensive or defensive techniques become adaptive, the environments that they reside in become non-stationary [27], putting further pressure on the IID foundations that ML builds on. This interaction can be approached more generally as a sequential zero-sum game [32], [25], [7]. To understand the implications of attacks and defenses becoming adaptive, we perform a theoretical investigation of their possible interactions with an ML-based system. As the agency that generates all the subsequent reasoning, we initiate our analysis from adversaries. In the following sections, the notation and terms we introduce are highlighted in red.

A. Attacks

The most compelling threat for deployed ML-based systems are hard-label, decision-based black-box attacks where no access is assumed to the model or its parameters, only the capacity to submit queries and receive discrete responses. One of the first decision-based attacks was Boundary Attack [8]. A large number of others followed – each inventive in its own way – that manage to improve the overall performance, typically measured as the lowest perturbation achieved for the minimal amount of queries submitted. Prominent examples are HSJA [14], Guessing Smart (BAGS) [11], Sign-Opt [16], Policy-driven (PDA) [52], QEBA [31], and SurFree [34].

White-box attacks like C&W [13] cannot function in black-box environments where there is no closed-form description of the inference pipeline. To facilitate optimization, decision-based attacks commonly initialize from a sample belonging to the target class, as it can be considered an adversarial example with an unacceptably large perturbation. This adjustment allows the task to be solved continuously, by minimizing the perturbation while always staying on the adversarial side of the boundary. Decision-based attacks share common aspects in their functioning that we can abstract through: given **starting** and **original** samples x_g and x_c respectively, the goal is to iteratively find adversarial **candidates** x_t, until the **distance** $\delta = d(x_t, x_c)$ is minimized. This process might follow different algorithmic approaches representing different geometrical intuitions; to describe them generally however, we can use a candidate generation policy:

$$\pi_\theta^{\mathcal{A}} = P\left(x_t | x_g, x_c, p^{\mathcal{A}}, s^{\mathcal{A}}\right) \tag{2}$$

that generates a candidate x_t, given x_g and x_c, with $p^{\mathcal{A}}$ the **parameters** and $s^{\mathcal{A}}$ the **state** of the attack. Considering that an attack episode evolves over time, and assuming the model always answers, this process can be construed as a Markov Decision Process (MDP) to be solved by a policy that minimizes d in the least amount of queries.

Consider a multinomial image classification model $\mathcal{M}$ under attack, based on a discriminant function $F : \mathbb{R}^d \to \mathbb{R}^m$, that for each input $x \in [0,1]^d$ generates an output $y := \{y \in [0,1]^m | \sum_{c=1}^m y_c = 1\}$ – a probability distribution over the m classes. By definition, black-box environments provide no access to these probabilities; instead one can only observe the result of the classifier C that returns the class with maximum probability:

$$C(x) := \arg\max_{c \in [m]} F_c(x) = D(F_c(x)) \tag{3}$$

with D being $\arg\max$, the decision function. The goal in targeted attacks is to change the **decision** $c_g \in [m]$ for a correctly classified example x, to a predefined **target** class $c_o \neq c_g$. This process can be facilitated through a function ψ which given a perturbed example x_t at step t, it returns a binary indicator of success:

$$\psi(x_t) = \begin{cases} 1 & \text{if} \quad C(x_t) = c_o \\ -1 & \text{if} \quad C(x_t) \neq c_o \end{cases} \tag{4}$$

As long as the model responds, ψ can always be evaluated, it thus constitutes the essential mechanism upon which decision-

based attacks build. The adversarial goal can then be described as the following constrained optimization problem:

$$\min_{x_t} d(x_t, x_c) \quad \text{s.t.} \quad \psi(x_t) = 1, \tag{5}$$

where the distance metric d is an ℓ_p-norm, with $p \in \{0, 1, 2, \inf\}$. As the threshold between adversarial and non-adversarial relies on the subjectivity of human perception, this optimization task highlights the imprecise nature of adversarial examples, something that is further aggravated in domains where visual affinity is of little importance. Customarily, successful or unsuccessful adversarial examples are delimited by an **threshold** ϵ on perturbation, where $d(x, x_t) \leq \epsilon$. While this constraint might not intuitively translate to non-visual domains, we nonetheless contend that minimal perturbation still remains a central property of adversarial examples.

The most relevant attacks in cybersecurity contexts are black-box and decision-based, and they are becoming increasingly effective. In HSJA for example, its optimization is guaranteed to converge to a stationary point of Eq. (5), and given typical ϵ values on perturbation imperceptibility this 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. Scalability is also an issue as typically in adversarial training a few steps (1-50) of a white-box attack – FGSM or PGD – are required, while decision-based attacks can take orders of magnitude more steps (queries) to produce an adversarial example.

Decision-based attacks typically search for the **optimal parameters** θ of generation policy (2), those that given x_c^i, with i denoting the i-th adversarial episode, minimize Eq. (5) in expectation:

$$\arg\min_{\theta} \mathbb{E}[\sum_{i=1}^{N} d(x_b^i, x_c^i)], \quad \text{s.t.} \quad \psi(x_b^i) = 1, \tag{6}$$

where x_b^i is the **best** adversarial example generated by policy π_θ^A during episode i. Given its dimensionality, it can be intractable to learn such a policy that modifies the input/problem space directly [37]; CIFAR-10, for instance, has more than 3K features to perturb. In AI enabled systems, the best practice is to freeze the model after validation so that no novel issues are introduced by retraining: for all queries x_t submitted during an attack session, we can assume that $F_0 = F_1 = ... = F_t, \forall t$. While this is representative of real-world settings, it is also what enables adversaries to discover adversarial examples that were not identified beforehand. The existence of adversaries which follow a candidate generation policy introduces however a behavior which can be observed and utilized by a defensive methodology. Consequently, while model-hardening approaches like adversarial training are *necessary*, they can be *insufficient* in defending against decision-based attacks.

Proposition III.1. *Let F_c be the discriminant function of the adversarially trained model $\mathcal{M}$. Then in order for $\mathbb{E}[\sum_{i=1}^{N} d(x_b^i, x_c^i)] \geq \epsilon$ in HSJA, two capabilities are necessary: a) a decision function $D' \neq \arg\max$, and b) additional context τ s.t for adversarial query x_t, $C(x_t) = D(F_c(x_t)) \neq D'(\tau, F_c(x_t))$.*

Intuitively, HSJA operates in 3 stages which repeat: a binary search that puts x_t on the decision boundary, a gradient estimation step, and projection step along the estimated gradient. If the model *always* responds truthfully, the adversary will be able to accurately perform all these steps and converge to the optimal adversarial; without loss of generality, we can extend this intuition to decision-based attacks which navigate the boundary. Secondly, the model should be able to differentiate between two, otherwise identical, queries when one is part of an attack and the other is not, and this is possible through a stateful representation; see Appendix A for the proof.

B. Defenses

Proposition III.1 shows that alternative classification policies are useful in the presence of decision-based attacks, for instance classification with rejection or intentional misdirection. In related work, rejection has been realized in the form of conformal prediction where model predictions are sets of classes including the empty one, or learning with rejection [6], [17]; while misdirection has emerged as a technique in adversarial RL and cybersecurity domains [22], [42]. While adversarially training the discriminant function F empirically shows the capacity to resist decision-based attacks, the manner in which the model responds has a 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 and adaptive defense can emerge. Active defenses have direct implications on attacks themselves however. Let us assume an agent carrying out an **active defense policy**:

$$\pi_\phi^{\mathcal{D}} = P(\alpha | x_t, s^{\mathcal{D}}),\ a \in \{0, 1\} \tag{7}$$

with x_t the query, $s^{\mathcal{D}}$ the **state** for the defense as created by past queries, and α the **binary decision**: when the query is deemed adversarial, it is rejected by returning $\alpha = 1$. If this policy is stationary the environment dynamics become in turn stationary and so, next to the adversarial task itself, bypassing the defense can *also* be formulated as an MDP to be solved. In two-player, zero-sum games, the moment an agent follows a stationary policy, it becomes *exploitable* through the reward obtained by an adversary [46]. Active defenses, as a consequence of decision-based attacks, entail adaptive adversaries.

Proposition III.2. *Against an active defense $\pi_\phi^{\mathcal{D}}$ and for time horizon T, a decision-based attack following a non-adaptive*

candidate generation policy $\pi_t = \pi_\theta^{\mathcal{A}}, \forall\, t \in [0, T]$ *will perform worse in expectation* (6)*, that is* $\mathbb{E}[\sum_{i=1}^{N} d(x_b^i, x_c^i)]^{\mathcal{D}} > \mathbb{E}[\sum_{i=1}^{N} d(x_b^i, x_c^i)]^{\not\mathcal{D}}$.

A proof for BAGS and HSJA is included in Appendix A. An adversary can reason, as a corollary to Proposition III.1, that such defenses *have to* be in place as it is suboptimal not too. However, there is a second reason to consider adaptive attacks even in the absence of active defenses, as attack policies are often not optimal out-of-the-box. Adapting attack policies can be seen as performing hyperparameter optimization and as an approach has proven very effective in other black-box or expensive-to-evaluate contexts, like Neural Architecture Search and Data Augmentation [53], [36]. The empirical results in Section V further indicate this correspondence between adaptive and self-optimizing, where the adaptive versions of attacks outperform the non-adaptive, more so against active defenses.

Consider now an active defense that is based on a similarity or a conformal metric. In the twofold meaning of adaptive we introduced in Section II, inventing control implies the *capability* to bypass a similarity based defense; adaptive control implies strategy instead, the active control of the available tools to evade the defense [1]. Then the adversary's task is to find the optimal adversarial policy that *also* evades rejection. The straightforward way to achieve this is by adapting policy (2) itself. Notably, and despite the discrete and black-box context, this optimization can be *fully* gradient-based [45]. We now demonstrate how adaptively controlling decision-based attacks recovers the **gradient-based** solvability of the black-box optimization task despite *neither* the active defense *nor* the model itself being accessible in closed-form.

Theorem III.1 (Adversarial Policy Gradient)**.** *Given model* $\mathcal{M}$ *with an active defense* $\pi_\phi^{\mathcal{D}}$ (7)*, adaptive candidate generation policy* $\pi_\theta^{\mathcal{A}}$ (2) *that generates episodes* τ *of queries* x_t*, and a reward* $r(\tau) = \sum_{x_t \in \tau} \neg\alpha$*, the optimal evasive policy* $\mathcal{E}$ *is obtained via the gradient of the policy's expected reward* $\mathbb{E}_{\pi_\theta^{\mathcal{A}}}[r(\tau)]$.

The proof is included in Appendix A. So far we have established that, **a)** in the presence of decision-based attacks, active defenses are necessary, yet conditional on adversarial agency they are insufficient and, **b)** by mere observation of discrete model decisions, adaptive attacks can become optimal in terms of both evasion and efficiency. The last piece of the puzzle is turning active defenses also adaptive.

Proposition III.3. *An active defense* $\pi_\phi^{\mathcal{D}}$ (7) *achieves its optimal, i.e. maximizes the expectation* $\mathbb{E}[\sum_{x_t \in \tau} P(\alpha | x_t, s)]$*, by adapting its policy against a stationary evasive policy* $\mathcal{E}$.

Proof. As offensive and defensive policies are strictly competitive, we can define the reward ρ of the defensive policy as $\rho(\tau) = \sum_{x_t \in \tau} \alpha$, then by making $\pi_\theta^{\mathcal{A}}$ stationary and $\pi_\phi^{\mathcal{D}}$ adaptive in Theorem III.1, we can reason that the optimal defensive policy is determined via the gradient of the expected reward $\mathbb{E}_{\pi_\phi^{\mathcal{D}}}[\rho(\tau)]$. □

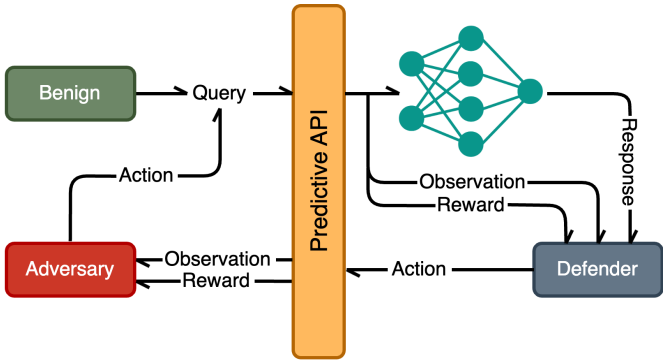

Fig. 2. Schematic model of an AMG environment. Due to the inherent uncertainty of behavior at either side of the interface, it is a partially observable RL environment mirrored for each agent where one's decisions become the other's observations.

C. Adversarial Markov Games

By reasoning on both offensive and defensive capabilities, we highlighted why we cannot consider them independently from each other. 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 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 this context as well as most cybersecurity environments. Model-free methods however can learn optimal offensive and defensive responses directly through interaction with the environment [42], [41], obviating the need to learn a model of it or to find *exact* solutions the to bi-level optimization task like Eq. (1) that is inherently NP-hard to solve [10].

To that end, Turn-Taking Partially-Observable Markov Games (TT-POMGs) introduced by Greenwald et al. [23] is a generalization of Extensive-Form Games (EFGs), widely used representations for non-cooperative, sequential decision-making games of imperfect and/or incomplete information; an apt formalism for decision-based attacks and defenses. Another nice property of TT-POMGs is that they can be transformed to equivalent belief state MDPs, significantly simplifying their solution.

The competition underlying adversarial example generation has been explored in no-box and white-box settings [7], [21]. We instead focus on decision-based, interactive environments which are assumed to have unknown but stationary dynamics: any agency present is considered part of the environment and therefore fixed in its behavior. By folding the strategies of other agents 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 not only from a theoretical perspective, but 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 modeling, computing the best response strategy in the simulated environment will correspond to the optimal defense.

This environment that encompasses adversarial attacks, adversarial defenses, and benign queries, can be construed as an Adversarial Markov Game (AMG) – a special case of TT-POMG – and is depicted in Figure 2. 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 goal of each player i is to determine a policy $\pi^i(A^i|O^i)$ that, given the policy of the other(s), maximizes their expected reward. When a player employs a stationary policy, the AMG reduces to a belief-state MDP where the other interacts with a fixed environment. The game is sequential and turn-taking, so each player i chooses an action a from their set of actions A^i which subsequently influences the observations of others.

We have shown that an adaptive defense policy $\pi_\phi^{\mathcal{D}}$ is necessary to deter decision-based attacks, and as a consequence the candidate generation policy $\pi_\theta^{\mathcal{A}}$ has to be adaptive in turn. As without implausible assumptions one cannot assume access to the exact state of the other agent, the states $O^{\mathcal{D}}$, $O^{\mathcal{A}}$ are partial observations of the complete state $\mathcal{S}$ of the full game. When human agents compete by holding beliefs about each other, they engage in recursive reasoning that in theory of mind is encountered as [I believe that [my opponent believes [that I believe...]]]. In the study of opponent modeling, considering other agent policies as stationary and part of the environment is equivalent to *0th* level recursive reasoning: 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], [51]. However, as AMGs can be solved by single-agent RL algorithms, we consider more involved recursive reasoning out of scope and perform the empirical evaluation without building explicit models of opponent behavior.

IV. Environment Specification

The empirical study we conduct in Section V is comprised diverse instantiations of the general theoretical framework introduced in Section III. Naturally, when working forward from the general to the particular, concrete design choices have to be made when specifying the latter, choices that have considerable influence on the results. To elucidate our proposed robustness evaluation methodology, in this section we provide the concrete details on the threat model and the environment.

Threat Model. Our AMG framework describes a two-player competitive game; while extensible to more players, in this work we assume that at a given moment only one attack takes place. From the defensive perspective, incoming queries can be either benign or part of an attack. An assumption that influences the effectiveness of stateful detection is that queries can be attributed to UIDs, e.g., an IP address or a user account. However, adversaries can collude, create multiple accounts, use VPNs, or in fact accounts and IP addresses might not even be required to query the model. To address this, we treat queries irrespective to their source: a strictly more challenging setting for stateful defenses where we operate solely on the content of queries and not on any other metadata, similar to [30]. Unlike Blacklight however, instead of rejecting queries, something that in itself can provide *more* information to the adversary and thus facilitate evasion [20], we misdirect by returning the second highest probability class. Furthermore, Gaussian noise is added to the benign queries to simulate a noisy channel and a shift in distribution so that is not trivial for a defense to tell adversarial noise apart. In summary, the black-box threat model we consider is delineated as follows:

does adversarial machine learning matter? we focus on black-boxn as the real in vivo threat model

- **Assets:** Trained and deployed model $\mathcal{M}$ with corresponding weights w.
- **Agents:** Adversary; defender; benign user.
- **Adversary Goal:** Generate minimal perturbation adversarial examples in as few queries as possible, while evading the defense.
- **Defender Goal:** Stop the adversary from generating adversarial examples, while preserving the correct functionality of the model $\mathcal{M}$ on benign users.
- **Adversary Knowledge:** The model $\mathcal{M}$ is known as the black-box function that transforms inputs $x \in [0, 1]^d$ to outputs $c \in [m]$, m being the number of classes. The weights w or the closed-form expression of $\mathcal{M}$ are unknown, as unknown is if an active defense $\pi_\phi^{\mathcal{D}}$ exists or not.
- **Defender Knowledge:** The defender observers the content of incoming queries only, without knowing if they come from a benign user or the adversary.
- **Adversary Capabilities:** Adapt the parameters of the attack and optionally any evasive transformations; in essence, adapt the candidate generation policy $\pi_\theta^{\mathcal{A}}$.
- **Defender Capabilities** For each query x, decide between answering truthfully with the true prediction $C(x) = c_0$ or misdirecting with the second highest probability class c_1.

Similarity. Decision-based attacks typically follow a policy that generates successive queries: these exhibit degrees of similarity which can be quantified by an appropriate l_p norm. If that norm is computed on the original inputs however, an adversary can *invent control* (see Fig. 1) by employing evasive transformations the model is invariant to and bypass the similarity detection. To account for this capability, we

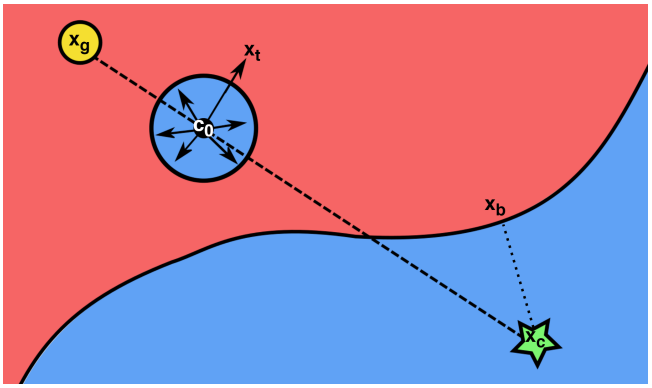

Fig. 3. Misdirection in a two-dimensional decision boundary. The adaptive defense controls a single parameter, the hypersphere radius around c_0 (the last known adversarial); for queries x_t that fall within this hypersphere the model responds with a non-adversarial decision. x_g is the starting sample, x_c the original, and x_b the best possible adversarial.

train a Siamese network with contrastive loss in order to learn a latent space $\mathcal{L}(\cdot)$ where similar inputs are mapped close together, unaffected by added noise or transformations on the inputs. For the stateful characterization of queries, we use two queues: one for the detected adversarial queries as determined by the defensive agent, and one for the benign or undetected ones.

Active defense. Recall that decision-based attacks evaluate a Boolean-valued function to determine if the query is adversarial or not; a straightforward counter to this behavior is to misdirect by returning a decision different from the actual through a system of confinement. When new query x_t is received, a state is constructed based on x_t and the queue $c_{-n}, c_{-n-1}, ..., c_0$ of known adversarial queries. Based on this state, the defensive agent takes a single continuous action $\alpha \in \mathbb{R} \mid 0 \leq \alpha \leq 1$ that is the radius α of a hypersphere around the last known adversarial query c_0 in the latent space $\mathcal{L}$. If $\|\mathcal{L}(x_t) - \mathcal{L}(c_0)\|_2 < a$ the query is considered adversarial and is appended to the adversarial queue as the latest c_0. The confinement system is depicted in Figure 3.

Adaptivity. No evaluation in AML is complete without considering adaptive adversaries; a notion we expand in this work, that is with the instruments to bypass the defense *and* their optimal configuration. As stateful defenses are so far similarity based, to bypass them intuition points towards input transformations the model is invariant to. For a given query x_t we want to compute a transformation $x'_t = T(x_t)$ so that $||x'_t - x_t||_2 \gg ||x_t - x_{t-1}||_2$ while $F(T(x_t)) \approx F(x_t)$. Depending on magnitude and composition of transformations T, the identity $F(T(x_t)) = F(x_t)$ might not always hold. As we also demonstrate in Section V, T interferes with the perturbations the adversarial policy generates itself: the performance and evasiveness of an attack are thus in a natural trade-off.

So what is the correct composition of transformations T to apply? When shall T be applied, and how does it affect the attack fundamentals? The transformations T can be considered as a set of additional controls, and like attack parameters they themselves can be suboptimal out-of-the-box [18]. Thus the combined control of attack and evasion parameters is a *prerequisite* to properly assess the strength of a defense. Their 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 precise optimal configuration and strategy of the attack.

Agents & Environments. Unlike common competitive games, in AMGs the two players have different action and state sets. AMGs are also asymmetric in the playing cadence: while the defender plays every round, the adversary might wait one to several rounds; HSJA for example is controlled on the iteration rather on the query level. Training is complicated further given that the experience upon which each agent learns arrives only *after* the opponent moves. We address these complications by developing custom learning environments for agents with asymmetries, with delayed experience collection, and asynchronous training, build with the OpenAI Gym and Stable Baselines 3 libraries [9], [40].

States & Actions. For the definition of the states we used and their rationale, we point the reader to Appendix B. For actions, we control BAGS through 4 parameters: orthogonal step size, source step size, mask bias, and Perlin bias. HSJA is controlled by 3: the gradient estimation radius, the number of estimation queries, and the jump step size. All evaluations start from controlling these attack parameters *only*; if the active defense proves impossible to defeat, we introduce additional knobs that control the magnitude and probability of transformations on the input, with the goal to evade detection *while* preserving semantic content and hence the correct classification. The range of transformations we experimented with as well as their magnitude and probability are listed in Table V. Finally, in both BAGS and HSJA the active defense consists of an 1-dimensional continuous action that controls the radius of confinement, as depicted in Fig 3.

Rewards. The success of any RL task relies heavily on *how* it is rewarded. Engineering an effective reward function is often non-trivial and has many intricacies, as reward hacking and specification gaming are common phenomena and the learned behavior can vary significantly [2]. For adversaries, the rewards we experimented with are variations on minimizing the distance to the original example – with extra reward shaping based on the fundamental operation of each attack – while defenders are rewarded or penalized for intercepting adversarial or benign queries respectively. The rewards in closed-form are included in Appendix B.

V. Evaluation

For evaluation, we define a range of scenarios intended reflect all possible and realistic combinations between adversarial attacks and defenses, and their adaptive versions. Concretely, the research questions we want to validate are: 1) Are active defenses a necessary complement to model hardening and to what extent? 2) Are attacks more threatening

when adaptive, i.e., do they outperform their vanilla versions *and* evade active detection? 3) If yes, can active defenses recoup their performance by also turning adaptive?

Metrics. We employ ASR (Attack Success Rate) and ℓ_2 norm of the perturbation. For the former we set a fixed threshold of 3 for consistency between experiments, while the latter is a more fine-grained metric well suited for comparing baseline attacks and defenses and their adaptive versions, as it is also not based on an arbitrary perceptual threshold that when changed yields widely varying results. The budgets we evaluate over are 1K, 2K and 5K queries. As robustness and classification accuracy are typically in trade-off, the third metric of interest is the accuracy on benign samples (Clean Acc.) that the original model and the active defense achieve together.

A. Evaluation Setup

The explicit goal for the agents is to learn offensive or defensive policies that are *general*: they transfer to *any* other evasion task. Thus after training and validating the agents, the final performance is reported on a fixed hold-out set of 100 adversarial episodes where the starting and original samples are selected at random. As is best practice in AML, candidate samples are only those that are correctly classified by the model. For each scenario we perform a limited hyperparameter and reward function exploration (max 30 trials), with the intention to root out poor combinations rather than exhaust the search space, described in more detail in Appendix D.

The black-box attacks we render adaptive and evaluate are **BAGS** and **HSJA**, as they represent two fundamentally different approaches, are highly effective, *and* have the highest evasion potential [30]. BAGS is a stochastic, search-based method where every query submitted is a new and potentially better adversarial example. Contrastively, HSJA is deterministic and composed of 3 different stages where the queries are generated in an aggregated manner: the vast majority of them are not candidate adversarial examples but means to approximate the gradient at the decision boundary.

In training and evaluation, the adversarial game is played as follows: the adversary starts by submitting a query, then the defender responds either *truthfully* (the true model prediction) or by *misdirecting* (the second highest probability class). Then the environment decides with chance p if the adversary moves next, otherwise a benign query is drawn. In either case, it is the defender's turn; in testing they are also oblivious to the nature of the query and know only the content. All experiments are performed with $p = 0.5$; we also evaluate adaptive defenses when no attack is present ($p = 0$) in subsection C-A.

The scenarios for all possible combinations of (non-) adaptive attacks and defenses are repeated over two different datasets – CIFAR10 and MNIST – and over two models with the same architecture but with different training regimes: one with adversarial training and one without. The transition from single-agent to multi-agent RL hides challenges however: we approach the AMG as a belief-state MDP (the requirement of knowing the exact opponent policies is relaxed) and use PPO [41] agents to discover optimal policies that will also constitute best responses for the full AMG [51]. However, learning independently of other agency breaks the theoretical guarantees of convergence [49] – like scenarios 7 & 8 where both agents learn simultaneously. The full list of scenarios is:

0) **VA-ND** – Vanilla Attack / No Defense: Baseline performance of attacks (BAGS & HSJA) out-of-the-box, without any active defense.
1) **AA-ND** – Adaptive Attack / No Defense: How more optimal is the adaptive version of an attack compared to the baseline.
2) **VA-VD** – Vanilla Attack / Vanilla Defense: The performance of our active defense, but the non-adaptive version with an empirically defined detection threshold.
3) **AA-VD** – Adaptive Attack / Vanilla Defense: Similar to scenario (2), but now the attack is adaptive.
4) **VA-AD** – Vanilla Attack / Adaptive Defense: The first scenario where the active defense is also adaptive, against the baseline adversary.
5) **AA-TD** – Adaptive Attack / Trained Defense: After the adaptive defense is optimized, its policy is fixed and an adaptive attack is trained against it.
6) **TA-AD** – Trained Attack / Adaptive Defense: The best policy found in the previous scenario is fixed and an adaptive defense is trained against it.
7) **AA-AD** – Adaptive Attack / Adaptive Defense: The first scenario where both agents learn simultaneously, making the environment non-stationary. In practice, the convergence will vary and depend on the chosen hyperparameters and rewards. Here we report the best-case for the attack.
8) **AA-AD** – Adaptive Attack / Adaptive Defense: Same scenario as before, but it reports the best-case for the defense instead.

For each successive scenario we evaluate with the most successful policy found, as this is best practice in Markov Games: the worst case opponent policy is fixed and then a counter to it is learned [32], [46]. Fixing other policies when computing a best response is representative of learning in cybersecurity environments that can contain a range of agents, with the additional benefit of converting the problem to single-agent that, as detailed in Section III-C, one can solve with RL.

SOTA Comparison. In Scenarios 0-8 we evaluate all possible combinations between our attack and our defense. As a baseline to compare to, we additionally evaluate our approach to the state-of-the-art stateful defenses and adaptive attacks, that is Blacklight [30] and OARS [20] respectively. We implement both Blacklight and OARS in our interactive multi-agent environments by using their publicly available code and parameters. To make a fair comparison with OARS, as our environments do not return a rejection signal, rejection coincides with non-adversarial decision. We thus define 5 additional scenarios:

9) **VA-BD** – Vanilla Attack / Blacklight Defense: Baseline performance of the attacks against Blacklight.
10) **OA-BD** – OARS Attack / Blacklight Defense: OARS against Blacklight.

TABLE II

ASR AND MEAN l_2 PERTURBATION FOR 1K, 2K, AND 5K QUERIES FOR CIFAR-10, AGAINST NORMALLY AND ADVERSARIALLY TRAINED MODELS. CLEAN ACC. REPORTS THE ACCURACY ON BENIGN QUERIES OF THE BASE MODEL PLUS ANY DEFENSES PRESENT; IN THE FIRST TWO SCENARIOS (NO ACTIVE DEFENSE) THE BASELINE CLEAN ACCURACY IS REPORTED. YELLOW SCENARIOS DENOTE THE BASELINE ATTACK PERFORMANCE, GREEN DENOTE ACTIVE AND/OR ADAPTIVE DEFENSES, AND RED DENOTE ADAPTIVE ATTACKS. THE ASTERISK DENOTES WHERE INPUT TRANSFORMATIONS WERE USED FOR EVASION.

Adv. Trained	Scenario	CIFAR-10 Gap: 20.01									
		BAGS				HSJA				Clean Acc.	
		1K	2K	5K	ASR	1K	2K	5K	ASR	BAGS	HSJA
✘	0: VA-ND	8.27	7.86	7.26	5%	3.42	1.43	0.41	100%	91.69	91.69
	1: AA-ND	1.26	0.71	0.49	100%	3.14	1.31	0.39	100%	91.69	91.69
	2: VA-VD	15.27	15.26	15.20	0%	11.14	10.81	10.33	7%	91.68	91.68
	3: AA-VD	2.63	2.03	1.77	93%	5.68	3.61	2.12	85%	91.69	91.69
	4: VA-AD	20.01	20.01	20.00	0%	17.17	16.35	15.56	0%	91.60	91.50
	*5: AA-TD	6.28	5.45	4.52	30%	13.19	11.82	10.69	2%	91.52	91.46
	*6: TA-AD	19.52	19.40	18.95	0%	16.48	16.13	15.69	0%	91.38	91.62
	*7: AA-AD	9.95	9.80	9.80	5%	10.30	9.04	7.55	23%	91.66	91.55
	*8: AA-AD	19.85	19.85	19.85	0%	14.46	13.93	13.08	1%	91.69	91.37
✔	0: VA-ND	8.72	8.42	7.94	4%	3.73	1.74	0.75	100%	87.76	87.76
	1: AA-ND	1.74	1.13	0.79	100%	3.64	1.77	0.73	100%	87.76	87.76
	2: VA-VD	15.42	15.35	15.20	0%	11.10	10.73	10.38	4%	87.72	87.73
	3: AA-VD	2.82	2.26	2.06	81%	5.66	3.36	1.94	86%	87.74	87.74
	4: VA-AD	20.01	20.01	20.00	0%	17.06	16.40	15.81	0%	87.66	87.66
	*5: AA-TD	8.48	7.68	6.82	9%	13.59	12.65	11.39	1%	87.58	87.52
	*6: TA-AD	19.58	19.40	18.95	0%	16.60	16.26	15.99	0%	87.50	87.68
	*7: AA-AD	10.43	10.24	10.17	1%	10.21	9.22	7.82	12%	87.73	87.61
	*8: AA-AD	19.86	19.86	19.86	0%	15.71	15.35	14.30	1%	87.67	87.40

11) **AA-BD** – Adaptive Attack / Blacklight Defense: Our adaptive attack against Blacklight.
12) **OA-TD** – OARS Attack / Trained Defense: OARS against our trained defense from Scenario 6.
13) **OA-AD** – OARS Attack / Adaptive Defense: Our adaptive defense retrained against OARS.

We run our experiments on multiple machines, however to give an idea of the model complexity of our defense, on an Intel i7-7700 CPU one forward pass in CIFAR – that is one response to the query – takes 700M FLOPs and 8 ± 1.4 ms.

B. Results

For consistency and comparability between evaluations, all results are from the *same* 100 test episodes. The **gap** value denotes the ℓ_2 perturbation that initially separates the starting and best adversarial example, averaged over the 100 episodes. By testing the trained agents on budgets higher than 5K we discovered that the trend in reducing ℓ_2 holds; to make the agent training tractable and the evaluation wider however, we limit the maximum query budget per adversarial episode to 5K. Tables II & III report the results for CIFAR10, while VIII reports MNIST. Overall, the empirical results help us extract and highlight several important insights, practical observations, and general implications for the broader AML field:

- When comparing the upper and lower halves of each table, we can observe that adversarial training adds a limited amount of robustness; otherwise, *the practical effect of adversarial training is a tax on the attacker*, forcing them to expend more queries for the same perturbation or having higher perturbation for the same query budget.
- Our adaptive defense (AD) outperforms both Blacklight (BD) and non-adaptive stateful (VD), also when transferred (S12). In HSJA, it reduces ASR by ∼90% when trained against OA specifically, while it offers similar protection to BD when transferred from another attack.
- Even against the strongest attacker and for the worst case (7), our AD keeps ASR as low as 23%.
- Our adaptive attack (AA) outperforms OARS (S11) and vanilla attacks (VA) by a wide margin and *without* access to rejection sampling and irrespective of the defense it faces; the only exception is our adaptive defense (AD), which it has very limited success against.
- Evasive transformations interfere with the attack operation, exemplified by the difference between BAGS and HSJA in (5); for the attack to reach its full potential, they have to be adaptively controlled together.
- The initial performance of an attack can be misleading: at first glance HSJA appears to be the better one but it is often outperformed by adaptive BAGS, especially in CIFAR and against active defenses.
- The advantage of AA is much more pronounced against active defenses, where they significantly outperform non-adaptive versions.
- The performance of both attacks deteriorates considerably against active defenses, however the latter reach their full potential only when *also adaptive*.
- We observe that between scenarios 1-8, where an agent trains against the best opponent policy as previously discov-

TABLE III
ASR AND MEAN l_2 PERTURBATION FOR CIFAR-10, COMPARING OUR ADAPTIVE ATTACK (AA) AND ADAPTIVE DEFENSE (AD) TO BLACKLIGHT (BD) AND OARS (OA).

Adv. Trained	Scenario	CIFAR-10 Gap: 20.01									
		BAGS				HSJA				Clean Acc.	
		1K	2K	5K	ASR	1K	2K	5K	ASR	BAGS	HSJA
✘	09: VA-BD	9.55	9.32	9.17	0%	8.41	8.19	7.80	15%	91.71	91.71
	10: OA-BD	9.46	9.46	9.46	1%	6.54	5.83	4.67	50%	91.71	91.71
	11: AA-BD	2.26	1.39	1.32	98%	4.55	3.08	2.44	78%	91.71	91.71
	12: OA-TD	20.01	20.01	20.01	0%	7.07	6.38	5.53	50%	91.61	91.59
	13: OA-AD	20.01	20.01	20.01	0%	11.03	11.00	10.95	5%	91.61	91.69
✔	09: VA-BD	9.75	9.56	9.46	0%	8.67	8.50	8.28	7%	87.76	87.76
	10: OA-BD	9.79	9.79	9.79	1%	5.77	4.53	3.26	72%	87.76	87.76
	11: AA-BD	5.59	4.04	2.55	79%	5.59	4.04	2.55	79%	87.76	87.76
	12: OA-TD	20.01	20.01	20.01	0%	6.44	5.49	4.38	65%	87.66	87.64
	13: OA-AD	20.01	20.01	20.01	0%	11.31	11.12	10.97	7%	87.66	87.74

ered, ASR oscillates as following a fixed policy enables the learning of an optimal counter to it, but eventually plateaus.

- Different attack fundamentals respond differently to different defenses; the gradient estimation part of HSJA is naturally disadvantaged against similarity detection, while the jump and binary steps are advantaged.
- When devising an adaptive defense for HSJA, it proved nearly impossible to engineer a state the agent can learn on by leveraging our knowledge of the attack and its geometric functioning. What did prove effective, however, was pure computation[2]: we learned an embedding for the state with Contrastive Learning from raw input [24]. These state space transfer exceedingly well to other attacks also, like BAGS.
- The first time an active defense effectively resists adaptive attacks is in (5) for CIFAR. As this is first scenario to do so, we employed evasive transformations from then onwards.

VI. DISCUSSION

Typically, the robustness adversarial training provides is against *all* adversarial examples under the same ℓ_p-norm; we do however discover that active defenses can transfer between attacks (Scenario 12) and as such they can be used jointly as complementary approaches. Our work has several implications for performing robust inference in the real-world. While adversarial training remains the most reliable defense, the amount of robustness it imparts will vary and even be insufficient. We demonstrated how AI-enabled systems are susceptible to adaptive adversaries that *devise* new evasive techniques and *control them jointly* with other attack parameters. This has been achieved in the *fully black-box* case and *against active defenses*. Even more concerning is the level of threat that such systems face from AI-enabled adversaries, as it is straightforward to generalize Theorem III.1 to any other domain or modality. This rekindles the proverbial arms race, where as a consequence defenses should also be backed by equally capable AI.

[2]This reminds us of Sutton's Bitter Lesson [44], the observation that progress in AI is often driven by gains in computation rather than problem-specific expert knowledge.

Limitations. To keep the amount of evaluations practical, we narrowed the scope to targeted attacks and to ℓ_2 as the more suitable norm for visual similarity. Targeted attacks are strictly more difficult to perform than untargeted, while in binary classification targeted and untargeted coincide; the latter is also the prevalent mode in cybersecurity contexts. Our framework, however, can accommodate any adversarial goal or metric.

A simplifying assumption we make is that only one attack can take place at a time; however, the queuing methodology we use for incoming queries is readily extensible to handle concurrent attacks. While we demonstrate how our stateful defense transfers between attacks, another possibility to explore is training the defense on multiple kinds of attacks. Finally, in our evaluation we focus on a wide range of adaptive and non-adaptive scenarios where agents learn concurrently, thus the number of datasets we experiment with is limited; our empirical study is backed by an extensive theoretical analysis however that supports the generality of our findings in any context.

Future Work. The AMG framework we introduce is general by design and can accommodate the learning of optimal offensive and defensive policies in any domain of interest beyond image classification. A promising path for future research is the extension of adaptive attacks and defenses to other domains and modalities, e.g. tasks like malware, bot, and network intrusion detection, precisely because our approach circumvents the obstacle of computing and mapping gradients to feasible perturbations and operates instead directly on the problem space [37]. Another compelling but formidable challenge is automating the adaptive evaluations in AML, that is adapting beyond a specification by inventing instruments to bypass defenses and thus imparting controllability to adversarial tasks. Finally, in our work we considered opponent agency as part of the environment; other domains, like malware analysis, might benefit from explicit opponent modeling.

TABLE IV

MEAN l_2 PERTURBATION FOR 1K, 2K, AND 5K QUERIES AND ACCURACY ON CLEAN DATA FOR MNIST. THE EVALUATION SCENARIOS ARE IDENTICAL TO TABLES II AND III.

Adv. Trained	Scenario	MNIST Gap = 10.62									
		BAGS				HSJA				Clean Acc.	
		1K	2K	5K	ASR	1K	2K	5K	ASR	BAGS	HSJA
✘	0: VA-ND	5.30	5.28	5.26	3%	3.59	3.07	2.61	73%	99.37	99.37
	1: AA-ND	2.74	2.57	2.47	78%	3.61	3.09	2.60	74%	99.37	99.37
	2: VA-VD	7.44	6.66	5.63	22%	5.82	5.78	5.73	2%	99.34	99.20
	3: AA-VD	3.79	3.66	3.44	29%	3.54	3.09	2.77	61%	99.37	99.31
	4: VA-AD	10.57	10.57	10.57	0%	10.05	10.05	10.05	0%	99.31	99.30
	5: AA-TD	3.57	3.29	3.14	39%	5.00	3.97	3.38	36%	99.32	98.84
	6: TA-AD	10.62	10.62	10.62	0%	10.23	10.23	10.18	0%	99.28	99.34
	7: AA-AD	4.89	4.89	4.86	8%	5.06	4.76	4.38	36%	99.31	99.35
	8: AA-AD	10.62	10.62	10.62	0%	10.21	10.21	10.21	0%	99.32	99.23
	09: VA-BD	10.62	10.62	10.62	0%	5.65	5.65	5.65	2%	99.37	99.37
	10: OA-BD	10.62	10.62	10.62	0%	4.53	4.00	3.15	46%	99.37	99.37
	11: AA-BD	3.83	3.69	3.60	17%	4.18	3.66	3.19	52%	99.37	99.37
	12: OA-TD	10.62	10.62	10.62	0%	10.21	10.20	10.20	0%	99.22	99.28
	13: OA-AD	10.62	10.62	10.62	0%	10.21	10.20	10.20	0%	99.32	99.28
✔	0: VA-ND	5.26	5.25	5.24	2%	4.61	4.04	3.41	30%	99.15	99.15
	1: AA-ND	3.28	3.08	2.96	51%	4.59	3.97	3.35	34%	99.15	99.15
	2: VA-VD	7.70	6.86	5.86	17%	5.81	5.78	5.76	2%	99.14	99.12
	3 AA-VD	4.18	4.08	3.86	22%	4.63	4.27	3.86	25%	99.13	99.15
	4: VA-AD	10.55	10.55	10.55	0%	10.02	10.02	10.02	0%	99.09	99.08
	5: AA-TD	4.04	3.74	3.54	27%	5.82	5.09	4.26	16%	99.11	98.78
	6: TA-AD	10.62	10.62	10.62	0%	10.20	10.20	10.20	0%	99.06	99.06
	7: AA-AD	5.59	5.56	5.56	5%	5.47	5.16	4.99	14%	99.09	99.13
	8: AA-AD	10.62	10.62	10.62	0%	10.12	10.12	10.12	0%	99.10	99.01
	09: VA-BD	10.62	10.62	10.62	0%	5.65	5.64	5.64	1%	99.15	99.15
	10: OA-BD	10.62	10.62	10.62	0%	5.18	4.80	4.11	17%	99.15	99.15
	11: AA-BD	4.31	4.07	3.96	13%	5.04	4.65	4.20	19%	99.15	99.15
	12: OA-TD	10.62	10.62	10.62	0%	10.26	10.26	10.26	0%	99.00	99.06
	13: OA-AD	10.62	10.62	10.62	0%	10.26	10.26	10.26	0%	99.10	99.06

VII. CONCLUSION

With adaptive, decision-based attacks becoming more pervasive in multiple domains, every AI-based decision-making process that exposes a queryable interface is inherently vulnerable. To aggravate matters, this vulnerability cannot be mitigated by employing model hardening approaches like adversarial training alone. To fully defend in the presence of such attacks, active *and* adaptive defenses are necessary, and we demonstrate how optimal defensive policies can be learned. However, the existence of such defenses elicits in turn adaptive attacks which are able to recover part of their original performance.

We perform a theoretical and empirical investigation of decision-based attacks and stateful defenses under a unified framework we name "Adversarial Markov Games" (AMG). 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. Furthermore, our adversarial policy gradient theorem indicates that any combination of adversarial goals, be it performance, stealthiness, or disruption, can be optimized in a gradient based manner, even in the *complete* black-box case and in *any* modality. As new attacks and defenses appear and get broken regularly, our evaluation methodology is generally applicable within the outlined scope by transforming any approaches in the existing arms-race to their self-adaptive versions, thus ensuring accurate and robust assessment of their performance.

The AMG framework we introduce helps us reason on and properly assess the vulnerabilities of AI-based systems, disentangling the inherently complex and non-stationary task of learning in the presence of competing agency. By modeling the latter as part of the environment, we can simplify this task by computing a best response to the observed behavior. This is an important outcome for cybersecurity domains: as long as proper threat modeling is carried out, one can readily employ RL algorithms in order to devise optimal defenses, but only after they devised optimal attacks too.

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] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[3] H. S. Anderson, A. Kharkar, B. Filar, D. Evans, and P. Roth. Learning to evade static pe machine learning malware models via reinforcement learning. *arXiv preprint arXiv:1801.08917*, 2018.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[10] 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.

[11] 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.

[12] 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.

[13] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] L. Demetrio, B. Biggio, G. Lagorio, F. Roli, and A. Armando. Functionality-preserving black-box optimization of adversarial windows malware. *IEEE Transactions on Information Forensics and Security*, 16:3469–3478, 2021.

[20] 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.

[21] X.-s. Gao, S. Liu, and L. Yu. Achieving optimal adversarial accuracy for adversarial deep learning using stackelberg games. *Acta Mathematica Scientia*, 42(6):2399–2418, 2022.

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

[23] 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.

[24] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[25] 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.

[26] 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.

[27] 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.

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

[29] M. Juuti, S. Szyller, S. Marchal, and N. Asokan. Prada: protecting against dnn model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 512–527. IEEE, 2019.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] A. Pal and R. Vidal. A game theoretic analysis of additive adversarial attacks and defenses. *Advances in Neural Information Processing Systems*, 33:1345–1355, 2020.

[36] H. Pham and Q. Le. Autodropout: Learning dropout patterns to regularize deep networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9351–9359, 2021.

[37] 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.

[38] M. Pintor, L. Demetrio, A. Sotgiu, A. Demontis, N. Carlini, B. Biggio, and F. Roli. Indicators of attack failure: Debugging and improving optimization of adversarial examples. *Advances in Neural Information Processing Systems*, 35:23063–23076, 2022.

[39] 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.

[40] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann. Stable baselines3, 2019.

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

[42] 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.

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

[44] R. Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13:12, 2019.

[45] 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.

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

[47] 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.

[48] I. Tsingenopoulos, A. M. Shafiei, L. Desmet, D. Preuveneers, and W. Joosen. Adaptive malware control: Decision-based attacks in the problem space of dynamic analysis. In *Proceedings of the 1st Workshop on Robust Malware Analysis*, pages 3–14, 2022.

[49] K. Tuyls and G. Weiss. Multiagent learning: Basics, challenges, and prospects. *Ai Magazine*, 33(3):41–41, 2012.

[50] 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.

[51] 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.

[52] 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.

[53] B. Zoph and Q. V. Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

APPENDIX A
PROOFS

For a more intuitive understanding of the proofs, we first provide a high-level description of the attack fundamentals. **BAGS** [11] performs a random walk along the boundary between the adversarial and the non-adversarial regions, by first taking a random step orthogonal to the original image direction, then a source step towards it. The randomness in the directions searched is reduced by utilizing Perlin noise and masks computed on the difference between starting and original samples. **HSJA** [14] operates in 3 stages: a binary search that places the current best adversarial on the decision boundary, an estimation step that computes the gradient at that point of the boundary, and projection step along the estimated gradient. These steps repeat until convergence.

[Proposition III.1]

Proof. The proof is constructed in two parts: first that an alternative decision function D is required, and subsequently that a stateful representation of a query x_t is also required. Let us denote by x_c, x_g, x_t the original (unperturbed), the starting, and the current sample at step t respectively. Given a target class $c_0 \in m$ we can define a function:

$$S_{x_c}(x_t) = F_{c_0}(x_t) - \max_{c \neq c_0}(F_c(x_t)) \tag{8}$$

HSJA operates in 3 stages that alternate until convergence or when the maximum query budget is reached: 1) A binary search between x_g and x_c that places x_t on the decision boundary between the classes. 2) A gradient estimation stage that approximates $\nabla S(x_t)$. 3) A "jump" step along the direction of the gradient $\nabla S(x_t)$. For practicality we reiterate Eq. 9 of [14] which denotes that the gradient direction is approximated via the following Monte Carlo estimate:

$$\widetilde{\nabla S_{x_c}}(x_t, \delta) = \frac{1}{B} \sum_{b=1}^{B} \phi_{x_c}(x_t + \delta u_b) u_b \tag{9}$$

where $\{u_b\}_{b=1}^{B}$ are i.i.d. draws from the uniform distribution over the d-dimensional sphere, δ is a small positive parameter, and ϕ_{x_c} is the Boolean-valued function that all steps of HSJA rely on:

$$\phi_{x_c}(x_t) = \text{sign}(S_{x_c}(x_t)) = \begin{cases} 1 & \text{if} \quad S(x_t) > 0, \\ -1 & \text{if} \quad S(x_t) \leq 0. \end{cases} \tag{10}$$

Given an initial sample x_c, in search of adversarial examples HSJA applies the following update function that sequentially updates x_c:

$$x_{t+1} = a_t x_c + (1 - a_t) \left\{ x_t + \xi_t \frac{\nabla S_{x_c}(x_t)}{\|\nabla S_{x_c}(x_t)\|_2} \right\} \tag{11}$$

where ξ_t is a positive step size and a_t is a line search parameter in $[0, 1]$ s.t. $S(x_{t+1}) = 0$, i.e. the next query lies on the boundary. Now let us assume that the decision function D is $\arg\max$, i.e. $D : \mathbb{R}^m \mapsto \mathbb{N}^m$, $C(x) = D(F_c(x)) = \arg\max F_c(x)$, then from Eq. 3 and Eq. 8 we have:

$$\begin{aligned} S_{x_c}(x_t) > 0 &\iff C(x_t) = c_0 \\ S_{x_c}(x_t) < 0 &\iff C(x_t) \neq c_0 \\ S_{x_c}(x_t) = 0 &\iff C(x_t) = \{c_0, a\},\ a \neq c_0 \\ \implies S_{x_c}(x_t) \leq 0 &\iff C(x_t) \neq c_0 \end{aligned} \tag{12}$$

Let us define the function $\mathcal{I}$ of two variables:

$$\mathcal{I}(a, b) := \begin{cases} 1 & \text{if} \quad a = b, \\ -1 & \text{if} \quad a \neq b. \end{cases} \tag{13}$$

We can rewrite Eq. 10 through and Eq. 13 and inequalities 12 as follows:

$$\phi(x_t) = \mathcal{I}(C(x_t), c_0) \tag{14}$$

Provided that the gradient estimation happens at the decision boundary where $S(x_t) = 0$, Theorem 2 of [14] guarantees that the estimation of stage 2 is an asymptotically unbiased direction of the true gradient:

$$\widetilde{\nabla S_{x_c}}(x_t, \delta) \approx \nabla S_{x_c}(x_t), \delta \to 0 \tag{15}$$

For $b_t = 1 - a_t$ and by plugging Eqs 14 & 15 in Eq. 9, and the result in 11, we get:

$$\begin{aligned} x_{t+1} &= a_t x_c + b_t \left\{ x_t + \xi_t \frac{\nabla S_{x_c}(x_t)}{\|\nabla S_{x_c}(x_t)\|_2} \right\} \\ &= a_t x_c + b_t \left\{ x_t + \xi_t \frac{\frac{1}{B} \sum_{b=1}^{B} \phi_{x_c}(x_t + \delta u_b) u_b}{\| \frac{1}{B} \sum_{b=1}^{B} \phi_{x_c}(x_t + \delta u_b) u_b \|_2} \right\} \\ &= a_t x_c + b_t \left\{ x_t + \xi_t \frac{\frac{1}{B} \sum_{b=1}^{B} \mathcal{I}(C(x_t + \delta u_b), c_0) u_b}{\| \frac{1}{B} \sum_{b=1}^{B} \mathcal{I}(C(x_t + \delta u_b), c_0) u_b \|_2} \right\} \end{aligned} \tag{16}$$

Recall that we assumed $C(x) = \arg\max F_c(x)$, then the updates in Eq. (16) are *guaranteed* by Theorem 1 of HSJA [14] to converge to a stationary point x_b of Eq. (5). Given a standard threshold ϵ on imperceptibility and over N adversarial episodes, this implies that $\mathbb{E}[\sum_{i=1}^{N} d(x_b^i, x_c^i)] < \epsilon$. As C is the only term in the expression (16) that is affected by the model, we reach a contradiction and that an alternative classifier C' is required, with $C(x) = D(F_c(x))$. As the discriminant function F_c cannot change without retraining, it follows that $D' \neq \arg\max$, so that for the adversarial example x_t misclassified as c_0, Eq. 4 can return -1:

$$\begin{aligned} C(x_t) = c_0 \Rightarrow \psi(x_t) = -1 \\ \therefore C(x_t) = \hat{c},\ \hat{c} = \{c_0, m \setminus c_0, \varnothing\} \end{aligned} \tag{17}$$

where the empty decision $\varnothing$ denotes rejection and $\{m \setminus c_0\}$ denotes misdirection, i.e. intentional misclassification.

Decision-based attacks initiate from examples x_t that belong to the target class c_0; while for $t = 0$, x_t is not adversarial yet, it still is *part of* an ongoing adversarial attack. To deter the attack, a perfect defender would have to misclassify/reject this example; yet if an identical but benign example x_n was submitted, classifier C should preserve its capacity to classify it correctly. For the identical examples $x_n = x_t$ we want $C(x_n) \neq C(x_t)$, something possible only with context. This context τ, based on the sequence of incoming queries

$x \in \{x_t, t = 0, 1, 2, ..., N\} \cup x_n$, can be provided as an additional argument to the decision function $D(\tau, F_c(x))$ s.t. while $x_n = x_t \implies D(\tau, F_c(x_n)) \neq D(\tau, F_c(x_t))$. □

[Theorem III.1]

Proof. Let H be a distance-based detector that based on a stateful representation $\mathcal{S}(\tau, x_t)$ decides if a query x_t is adversarial or not, and we assume the candidate generation policy $\pi_\theta^{\mathcal{A}}$ can influence this representation through intrinsic parameters like source step size, or applying transformations the underlying model is invariant to (V). For a predefined threshold δ, if $\mathcal{S}(\tau, x_t) < \delta$ then $H(x_t) = 1$ indicating that x_t is part of an adversarial attack, otherwise $H(x_t) = 0$. Now for a number N of attack episodes of fixed length L and a reward function $r(\tau) = \sum_{t=1}^{L}(1 - H(x_t))$, the objective to be maximized is the expected reward of the adversarial policy:

$$\mathcal{J}(\theta) = \mathbb{E}_{\pi_\theta^{\mathcal{A}}}[r(\tau)] = \frac{1}{N}\sum_{i=1}^{N} r(\tau) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{L}(1 - H(x_t)) \tag{18}$$

As it is not possible to compute the derivative of an expectation, by using the Policy Gradient Theorem [45] we can transform it to the expectation of the product of the reward and the gradient of the logarithm of the policy:

$$\begin{aligned}\nabla\mathbb{E}_{\pi_\theta^{\mathcal{A}}}[r(\tau)] &= \mathbb{E}_{\pi_\theta^{\mathcal{A}}}[r(\tau)\nabla \log \pi_\theta(\tau)] \\ &= \mathbb{E}_{\pi_\theta^{\mathcal{A}}}[\sum_{t=1}^{L}(1 - H(x_t))\nabla \log \pi_\theta^{\mathcal{A}}(\tau)] \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{L}(1 - H(x_t))\nabla \log \pi_\theta^{\mathcal{A}}(\tau)\end{aligned} \tag{19}$$

To maximize this expectation empirically we sample e episodes from the environment, compute the gradient of the policy $\nabla \log \pi_\theta^{\mathcal{A}}(\tau)$ and then update the policy with a learning step a, $\theta' = \theta + \alpha \cdot \sum_0^e r(\tau)\nabla \log \pi_\theta^{\mathcal{A}}(\tau)$. After this process has converged through gradient ascent, and as the maximum in Eq. 18 is attained for $\sum_{i=1}^{N}\sum_{t=1}^{L} H(x_t) = 0$, the obtained policy $\pi_\theta^{\mathcal{A}}$ will correspond to the optimally evasive one $\mathcal{E}$. □

[Proposition III.2] We annotate terms with $\mathcal{D}$ when an active defense $\pi_\phi^{\mathcal{D}}$ is present, and with $\not{\mathcal{D}}$ otherwise.

Proof. **BAGS**. This attack is in effect a gradual interpolation from x_g towards x_c, by first taking orthogonal steps x_s on the hypersphere around x_c and then source steps towards x_c in order to minimize $d(x_c - x_b)$, where x_b is the best adversarial example found so far. The source step parameter $\epsilon = (1.3 - \min(\lambda_n, 1)) \cdot c$ — with λ_n the ratio of the n last queries x_t that are adversarial and c a positive constant — controls the projection towards x_c:

$$x_t = x_s + \epsilon \cdot (x_c - x_s) \tag{20}$$

Then if we again assume that a non-zero amount of the adversarial queries x_t is flagged as such by the defense, it follows that $\lambda_n^{\not{\mathcal{D}}} > \lambda_n^{\mathcal{D}}$ and from the definition of ϵ we get $\epsilon^{\mathcal{D}} < \epsilon^{\not{\mathcal{D}}}$. At given t, from Eq. (20) we get that $d(x_c, x_t^{\mathcal{D}}) > d(x_c, x_t^{\not{\mathcal{D}}})$, and ceteris paribus the expectation (6) will be larger with $\pi_\phi^{\mathcal{D}}$ present than without.

HSJA. We denote the queries during gradient estimation as $x_n = x_t + \delta u$, $u \sim Uniform_{Sphere}(d)$, the ratio of those x_n detected as adversarial by the active defense as $\eta \in [0, 1]$, and the estimate $\widetilde{\nabla S_{x_c}}(x_t, \delta)$ as u_t. We investigate the behavior of active defenses as the ratio of detections η goes to 1.

For $\eta = 1 \implies \mathbb{E}[\phi_{x_c}(x_n)] = -1$, and as u_b are uniformly distributed, from Eq. (9) we get:

$$\lim_{\eta\to 1} u_t = \lim_{\eta\to 1}\frac{1}{B}\sum_{b=1}^{B}\phi_{x_c}(x_t + \delta u_b)u_b = \frac{1}{B}\sum_{b=1}^{B} -u_b \tag{21}$$

At the limit of detection we observe that the gradient estimate u_t behaves like a uniformly drawn vector around x_t of shrinking size. By the Law of Large Numbers, as B increases the average direction of u_t will align with the expected value: that is a random direction on the unit hypersphere. However, due to the $\frac{1}{B}$ term, the size of u_t goes to 0. From Eq. (21) then we get: $\lim_{\eta\to 1} u_t = 0$. The gradient estimation step is followed by the "jump" step that computes x_{t+1} as follows:

$$x_{t+1} = x_t + \xi u_t \tag{22}$$

As the ratio of detections η approaches 1, we observe that the adversarial iterates x_{t+1} converge prematurely: then all else being equal and for given t, $d(x_c, x_t^{\mathcal{D}}) > d(x_c, x_t^{\not{\mathcal{D}}})$.

□

Appendix B
On States & Rewards

[States]. In POMDPs a single observation cannot constitute a Markovian state as typically there is some of dependence on the state history. Partial observability is handled by either using recurrent architectures for the policy and value networks, or by engineering a state that includes past information: in this work we opt for the later.

For BAGS, the adversary uses an 8-dimensional state representation with the following information normalized in the range $[0, 1]$: current amount of queries i, average queries that are adversarial a, the initial gap g, the current gap d, the location $l = \frac{d}{g}$, the slope $s = m - l$ where m is a moving average of the location, the frequency of improvement f, and r which is a moving average of the perturbation reduction n, In HSJA the state representation is slightly different: $r = \frac{n}{g}$, and $f = \frac{1}{j}$ with j number of jump steps in last iteration.

Regarding the defense, for HSJA (and to a lesser extent BAGS) it has been difficult to engineer a representative state for policies to effectively learn on. The knowledge of the attack internals and fundamentals, geometric properties and distances, model activations and logits, and any combination thereof, did not suffice. Ultimately we turned to pure computation to learn a state representation for the defensive agent. This representation is a 64-dimensional embedding of a CNN trained with triplet loss, on data generated by HSJA and benign queries, where the input is a tensor where the last query is

subtracted from the 25 most recent adversarial queries and then stacked together.

[Rewards].

The concrete definitions of the rewards for each type of agent are:

- BAGS adversary: with $x \in [1, 50]$ the number of queries to a better adversarial example and t the maximum queries: **R1** $= \frac{n \cdot x}{g}$ if $n > 0$ else 0 | **R2** $= \frac{n}{g \cdot (x+1)}$ if $n > 0$ else 0 | **R3** $= (1 - \sqrt{\frac{d}{g}})^2 - (1 - \sqrt{\frac{d+n}{g}})^2$ | **R4** $= \sqrt{i} \cdot R2$ | **R5** $= |\log(d/g)|$ if $i \geq t$ else 0 | **R6** $= \sqrt[4]{i} \cdot a$ | **R7** $= R4 + R6$.
- HSJA adversary: with e the gradient estimation steps: **R1** $= 2 \cdot n$ | **R2** $= \frac{-e}{1000} + R1$ | **R3** $= \frac{10 \cdot n}{d}$ | **R4** $= \frac{1}{d}$ | **R5** $= \frac{2 \cdot (g-d)}{g}$ if $i \geq t$ else 0 | **R6** $= 2 \cdot (0.5 - |\frac{a+1}{2} - 0.5|) + b$, where $b = \frac{j}{20}$ if $j < 3$ else 0 | **R7** $= R3 + R6$ — **R8** $= R5 + R6$.
- BAGS defender: where x_g is the starting sample, x_t the last query, x_b the best adversarial so far, s_t the average step size between queries, $h \in [0, 1]$ the last action of the defender, $z \in [0, 1]$ the ℓ_2 distance of x_t and the last known adversarial query in embedding space, x: **R1** $= |\log(0.1g + \|x_g, x_b\|_{\ell_2})| \cdot 0.1$ | **R2** $= |\log_{10} s_t|$ | **R3** $= \frac{g}{\|x_g, x_t\|}$ | **R4** $= -\psi(x_t)$, where ψ is Eq. 4 | **R5** $= h - z$.
- HSJA defender: where x_{BS} are queries during the binary search: **R1** $= 1 - 2(\frac{\|x_g, x_b\|}{g})$ | **R2** $= h - z$ | **R3** $= R2 - 2\psi(x_{BS})$ | **R4** $= -\|\psi(x_{BS})\|$ | **R5** $= R2$ if $\psi(x_t)$ else $2 \cdot R2$.
- For both BAGS and HSJA defenders, the aforementioned are the rewards when x_t is adversarial; when it is benign, the reward is $R = 1 - h$ if the model responded correctly, otherwise $R = -1$.

TABLE V
INPUT TRANSFORMATIONS.

Input Transformations	Magnitude	Probability
Brightness & Contrast	0 – 0.5	0 – 1
Random Horizontal Flip	–	0 – 1
Random Vertical Flip	–	0 – 1
Sharpness	0.8 – 1.8	0 – 1
Perspective	0.25 – 0.5	0 – 1
Rotation	°0 – °180	0 – 1
Uniform Pixel Scale	0.8 – 1.2	0 – 1
Crop & Resize	0.6 – 1	0 – 1
Translation	-0.2 – 0.2	0 – 1

APPENDIX C
ADDITIONAL EXPERIMENTS

Besides CIFAR-10, We followed the same evaluation protocol to assess all the scenarios mentioned in Section V also on the MNIST dataset; the results are shown in Table VIII.

A. Base Rate of Attacks

In all the evaluations so far we use a fixed probability $P(adv) = 0.5$ that an incoming query is adversarial. To assess how our adaptive stateful defenses (Scenarios 4 & 6) perform in the complete absence of attacks, we evaluate them with $P(adv) = 0$ *without* retraining; the results are shown in Table VI. We observe a small reduction in the accuracy on clean samples that can be attributed to the considerably different base rate of adversarial and benign queries. Note however that as the probability of adversarial queries is an intrinsic property of each environment, if the base rate of attacks changes the defensive agents can be retrained to adjust to it.

TABLE VI
CLEAN ACCURACY ON CIFAR-10 FOR SCENARIOS 4 & 6, WHERE $P(adv)$ DENOTES THE PROBABILITY THAT A QUERY IS PART OF AN ATTACK.

Adv. Trained	$P(adv)$	BAGS4	BAGS6	HSJA4	HSJA6
✘	0.5	91.55	91.38	91.59	91.62
	0.0	90.91	90.95	90.48	90.86
✔	0.5	87.61	87.50	87.58	87.68
	0.0	87.02	87.11	86.70	86.98

APPENDIX D
MODELS & HYPERPARAMETERS

The image classification models we use are ResNet-20 for CIFAR-10 and a standard 2 convolutional / 2 fully-connected layer NN for MNIST. For adversarially training models, we follow the canonical approach as described in [50]: the model is trained for 20 epochs, where the first 10 are trained normally and the last 10 on batches containing additional adversarial examples generated with 40 steps of PGD. For learning the similarity space, that is the metric space where defensive agents control the radius of interception around which a query is adversarial or not, we use a Siamese CNN. This network is trained with contrastive loss, where dissimilar examples are generated by adding Gaussian noise and performing evasive transformations on the input from the list in Table V. For the PPO agents trained for each scenario, we use the open source library Stable-Baselines3 [3]. Policies are parameterized by a two fully-connected layer NN; the hyperparameter search space is shown in Table VII.

A. Blacklight & OARS

For evaluating Blacklight [30] and OARS [20] we use their default hyperparameters without tuning them, as those are provided in the publicly available implementations. In particular, as OARS spends 200 extra queries per episode to adapt the proposal distribution, we add those to the evaluation budget. Additionally, as our defense is not rejection based, we replace the rejection decision with a non-adversarial one.

[3]https://github.com/DLR-RM/stable-baselines3

TABLE VII
HYPERPARAMETER SPACE FOR THE PPO AGENT.

Hyperparameter	**BAGS**	**HSJA**
learning rate	3e-3 – 1e-4	3e-3 – 1e-4
episode steps	600 – 3000	1000 – 5000
total steps	1e5 – 4e5	2e4 – 2e5
batch size	32 – 128	32 – 64
buffer	2048 – 2048	64 – 1024
epochs	20 – 20	20 – 20
gamma	0.85 – 0.99	0.9 – 0.99

TABLE VIII

MEAN l_2 PERTURBATION FOR 1K, 2K, AND 5K QUERIES AND ACCURACY ON CLEAN DATA FOR MNIST. THE EVALUATION SCENARIOS ARE IDENTICAL TO TABLES II AND III.

Adv. Trained	Scenario	**MNIST Gap = 10.62** BAGS 1K	BAGS 2K	BAGS 5K	BAGS ASR	HSJA 1K	HSJA 2K	HSJA 5K	HSJA ASR	Clean Acc. BAGS	Clean Acc. HSJA
✘	0: VA-ND	5.30	5.28	5.26	3%	3.59	3.07	2.61	73%	99.37	99.37
	1: AA-ND	2.74	2.57	2.47	78%	3.61	3.09	2.60	74%	99.37	99.37
	2: VA-VD	7.44	6.66	5.63	22%	5.82	5.78	5.73	2%	99.34	99.20
	3: AA-VD	3.79	3.66	3.44	29%	3.54	3.09	2.77	61%	99.37	99.31
	4: VA-AD	10.57	10.57	10.57	0%	10.05	10.05	10.05	0%	99.31	99.30
	5: AA-TD	3.57	3.29	3.14	39%	5.00	3.97	3.38	36%	99.32	98.84
	6: TA-AD	10.62	10.62	10.62	0%	10.23	10.23	10.18	0%	99.28	99.34
	7: AA-AD	4.89	4.89	4.86	8%	5.06	4.76	4.38	36%	99.31	99.35
	8: AA-AD	10.62	10.62	10.62	0%	10.21	10.21	10.21	0%	99.32	99.23
	09: VA-BD	10.62	10.62	10.62	0%	5.65	5.65	5.65	2%	99.37	99.37
	10: OA-BD	10.62	10.62	10.62	0%	4.53	4.00	3.15	46%	99.37	99.37
	11: AA-BD	3.83	3.69	3.60	17%	4.18	3.66	3.19	52%	99.37	99.37
	12: OA-TD	10.62	10.62	10.62	0%	10.21	10.20	10.20	0%	99.22	99.28
	13: OA-AD	10.62	10.62	10.62	0%	10.21	10.20	10.20	0%	99.32	99.28
✔	0: VA-ND	5.26	5.25	5.24	2%	4.61	4.04	3.41	30%	99.15	99.15
	1: AA-ND	3.28	3.08	2.96	51%	4.59	3.97	3.35	34%	99.15	99.15
	2: VA-VD	7.70	6.86	5.86	17%	5.81	5.78	5.76	2%	99.14	99.12
	3 AA-VD	4.18	4.08	3.86	22%	4.63	4.27	3.86	25%	99.13	99.15
	4: VA-AD	10.55	10.55	10.55	0%	10.02	10.02	10.02	0%	99.09	99.08
	5: AA-TD	4.04	3.74	3.54	27%	5.82	5.09	4.26	16%	99.11	98.78
	6: TA-AD	10.62	10.62	10.62	0%	10.20	10.20	10.20	0%	99.06	99.06
	7: AA-AD	5.59	5.56	5.56	5%	5.47	5.16	4.99	14%	99.09	99.13
	8: AA-AD	10.62	10.62	10.62	0%	10.12	10.12	10.12	0%	99.10	99.01
	09: VA-BD	10.62	10.62	10.62	0%	5.65	5.64	5.64	1%	99.15	99.15
	10: OA-BD	10.62	10.62	10.62	0%	5.18	4.80	4.11	17%	99.15	99.15
	11: AA-BD	4.31	4.07	3.96	13%	5.04	4.65	4.20	19%	99.15	99.15
	12: OA-TD	10.62	10.62	10.62	0%	10.26	10.26	10.26	0%	99.00	99.06
	13: OA-AD	10.62	10.62	10.62	0%	10.26	10.26	10.26	0%	99.10	99.06