Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning

Battista Biggio[a,b,*], Fabio Roli[a,b]

[a]*Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi 09123 Cagliari, Italy*
[b]*Pluribus One, Cagliari, Italy*

Abstract

Learning-based pattern classifiers, including deep networks, have demonstrated impressive performance in several application domains, ranging from computer vision to computer security. However, it has also been shown that adversarial input perturbations carefully crafted either at training or at test time can easily subvert their predictions. The vulnerability of machine learning to adversarial inputs (also known as adversarial examples), along with the design of suitable countermeasures, have been investigated in the research field of adversarial machine learning. In this work, we provide a thorough overview of the evolution of this interdisciplinary research area over the last ten years, starting from pioneering, earlier work up to more recent work aimed at understanding the security properties of deep learning algorithms, in the context of different applications. We report interesting connections between these apparently-different lines of work, highlighting common misconceptions related to the evaluation of the security of machine-learning algorithms. We finally discuss the main limitations of current work, along with the corresponding future research challenges towards the design of more secure learning algorithms.

Keywords: Adversarial Machine Learning; Evasion Attacks; Poisoning Attacks; Adversarial Examples; Secure Learning; Deep Learning

1. Introduction

Modern technologies based on pattern recognition, machine learning and, more generally, data-driven artificial intelligence (AI), especially after the advent of deep learning, have reported impressive performance in a variety of application domains, including speech and object recognition, and spam and malware detection. It has been thus surprising to see that they can easily be fooled by *adversarial examples*, i.e., carefully-perturbed input samples aimed to mislead detection at test time. This has brought considerable attention since 2014, when Szegedy et al. [1] and subsequent work [2–4] showed that deep convolutional neural networks for object recognition can be fooled by input images perturbed in a visually-indistinguishable manner.

Since then, an ever-increasing number of research papers have started proposing countermeasures to mitigate this threat, not only in the area of computer vision [5–11].[1] This huge and growing body of work has clearly fueled also a renewed interest in the research field known as *adversarial machine learning*, while also raising a number of misconceptions on how the security properties of learning algorithms should be evaluated and understood.

The primary misconception is about the start date of the field of *adversarial machine learning*, which is not 2014. This wrong start date is implicitly acknowledged in a growing number of recent papers in the area of computer security [5, 7, 10, 12–16] and computer vision [8, 9, 17], which only consider recent work on the security of deep networks against adversarial examples, almost completely ignoring previous work in adversarial machine learning.

To the best of our knowledge, the very first, seminal work in the area of adversarial machine learning dates back to 2004. At that time, Dalvi et al. [18], and immediately later Lowd and Meek [19, 20] studied the problem in the context of spam filtering, showing that linear classifiers could be easily tricked by few carefully-crafted changes in the content of spam emails, without significantly affecting the readability of the spam message. These were indeed the first adversarial examples against linear classifiers for spam filtering. In 2006, in their famous paper, Barreno et al. [21] questioned the suitability of machine learning in adversarial settings from a broader perspective, categorizing attacks against machine-learning algorithms both at training and at test time, and envisioning potential countermeasures to mitigate such threats. Since then, and independently from the discovery of *adversarial examples* against deep networks [1], a large amount of work has been done to: (*i*) develop attacks against machine learning, both at training time (poisoning attacks) [22–31] and at test time (evasion attacks) [18–20, 28, 32–37]; (*ii*) propose systematic methodologies for security evaluation of learning algorithms against such attacks [37–41]; and (*iii*) design suitable defense mechanisms to mitigate these threats [11, 18, 33, 42–46].

*Corresponding author
Email addresses: battista.biggio@diee.unica.it (Battista Biggio), roli@diee.unica.it (Fabio Roli)

[1] More than 150 papers on this subject were published on ArXiv only in the last two years.

The fact that *adversarial machine learning* was well-established before 2014 is also witnessed by a number of related events, including the 2007 NIPS Workshop on Machine Learning in Adversarial Environments for Computer Security [47], along with the subsequent special issue on the journal *Machine Learning* [48], the 2013 Dagstuhl Perspectives Workshop on Machine Learning Methods for Computer Security [49] and, most importantly, the Workshop on Artificial Intelligence and Security (AISec), which reached its 10th edition in 2017 [50]. Worth remarking, a book has also been recently published on this subject [51].

In this work, we aim to provide a thorough overview of the evolution of this interdisciplinary research area over the last ten years, from pioneering work to more recent work on the security properties of deep learning algorithms, in the context of different applications. Our goal is to *connect the dots* between these apparently-different lines of work, while also highlighting common misconceptions related to the security evaluation of machine-learning algorithms.

We first review the notion of arms race in computer security, advocating for a proactive security-by-design cycle that explicitly accounts for the presence of the attacker in the loop (Sect. 2). Our narrative of the security of machine learning then follows three metaphors, referred to as the *three golden rules* in the following: (*i*) know your adversary, (*ii*) be proactive; and (*iii*) protect yourself. Knowing the attacker amounts to modeling threats against the learning-based system under design. To this end, we review a comprehensive threat model which allows one to envision and simulate attacks against the system under design, to thoroughly assess its security properties under well-defined attack scenarios (Sect. 3). We then discuss how to proactively simulate test-time evasion and training-time poisoning attacks against the system under design (Sect. 4), and how to protect it with different defense mechanisms (Sect. 5). We finally discuss the main limitations of current work and the future research challenges towards the design of more secure learning algorithms (Sect. 6).

2. Arms Race and Security by Design

Security is an arms race, and the security of machine learning and pattern recognition systems is not an exception to this [39, 40]. To better understand this phenomenon, consider that, since the 90s, computer viruses and, more generally, Internet scams have increased not only in terms of absolute numbers, but also in terms of variability and sophistication, in response to the growing complexity of defense systems. Automatic tools for designing novel variants of attacks have been developed, making large-scale automatization of stealthier attacks practical also for non-skilled attackers. A very clear example of this is provided by *phishing kits*, which automatically compromise legitimate (vulnerable) websites in the wild, and hide phishing webpages within them [52, 53]. The sophistication and proliferation of such attack vectors, malware and other threats is strongly motivated by a flourishing underground economy, which enables easy monetization after attack. To tackle the increasing complexity of modern attacks, and favor the detection

Figure 1: Examples of *clean* (top) and *obfuscated* (bottom) spam images [54].

of never-before-seen ones, machine learning and pattern recognition techniques have been widely adopted over the last decade also in a variety of security application domains. However, as we will see throughout this paper, machine learning and pattern recognition techniques turned out not to be the definitive answer to such threats. They introduce specific vulnerabilities that skilled attackers can exploit to compromise the whole system, i.e., machine learning itself can be the *weakest link* in the security chain.

To further clarify how the aforementioned arms race typically evolves, along with the notions of reactive and proactive security, we briefly summarize in the following an exemplary case in spam filtering.

The spam arms race. Spam emails typically convey the spam message in textual format. Rule-based filtering and text classifiers are indeed used to classify emails as legitimate (ham) or spam. Spammers attempt to mislead these defenses by obfuscating the content of spam emails to evade detection, e.g., by misspelling *bad* words (i.e., words likely to appear in spam but not in legitimate emails), and adding *good* words (i.e., words typically occurring in legitimate emails, randomly guessed from a reference vocabulary) [20, 32, 42]. In 2005, spammers invented a new trick to evade textual-based analysis, referred to as *image-based spam* (or image spam, for short) [54, 55]. The idea is simply to embed the spam message within an attached image (Fig. 1, left). Due to the large amount of image spam sent in 2006 and 2007, countermeasures were promptly developed based on signatures of known spam images (through hashing), and on extracting text from suspect images with OCR tools [56]. To evade these defenses, spammers started obfuscating images with random noise patterns (Fig. 1, right). Ironically, this trick exploits similar techniques to those used in CAPTCHAs to protect web sites from spam bots. Learning-based approaches based on low-level visual features were then devised to discriminate between spam and legitimate images. Image spam volumes have since declined, but spammers have been constantly developing novel tricks to evade detection.

Reactive and proactive security. As discussed for spam filtering, security problems are often cast as a *reactive* arms race, in which the system designer and the attacker aim to achieve their goals by adapting their behavior in response to that of the opponent, i.e., *learning from the past*. This can be mod-

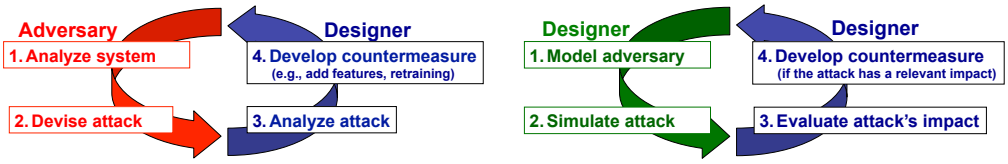

Figure 2: A conceptual representation of the reactive (left) and proactive (right) arms races for pattern recognition and machine learning systems in computer security [39, 40].

eled according to the following steps (Fig. 2, left) [39, 40]: (*i*) the attacker analyzes the defense system and crafts an attack to violate its security; and (*ii*) the system designer analyzes the newly-deployed attacks and designs novel countermeasures against them. However, *reactive* approaches are clearly not able to prevent the risk of *never-before-seen* attacks. To this end, the designer should follow a *proactive* approach to anticipate the attacker by (*i*) identifying relevant threats against the system under design and simulating the corresponding attacks, (*ii*) devising suitable countermeasures (if retained necessary), and (*iii*) repeating this process *before* system deployment (Fig. 2, right). In practice, these steps are facilitated by leveraging a thorough model of the attacker, as that discussed in the next section, which helps envisioning and analyzing a number of potential attack scenarios against learning-based systems.

3. Know Your Adversary: Modeling Threats

> *"If you know the enemy and know yourself, you need not fear the result of a hundred battles."* (Sun Tzu, The Art of War, 500 BC)

We discuss here the *first golden rule* of the proactive security cycle discussed in the previous section, i.e., how to model threats against learning-based systems and thoroughly evaluate their security against the corresponding attacks. To this end, we exploit a framework based on the popular attack taxonomy proposed in [21, 38, 57] and subsequently extended in [6, 28, 31, 39, 40], which enables one to envision different attack scenarios against learning algorithms and deep networks, and to implement the corresponding attack strategies. Notably, these attacks include training-time poisoning and test-time evasion attacks (also recently referred to as adversarial training and test examples) [2–4, 13–15, 24, 27, 29–31, 36, 39, 57]. It consists of defining the attacker's goal, knowledge of the targeted system, and capability of manipulating the input data, to subsequently define an optimization problem corresponding to the optimal attack strategy. The solution to this problem provides a way to manipulate input data to achieve the attacker's goal. While this framework only considers attacks against *supervised* learning algorithms, we refer the reader to similar threat models to evaluate the security of clustering [58–60], and feature selection algorithms [27, 61] under different attack settings.

Notation. In the following, we denote the sample and label spaces with $\mathcal{X}$ and $\mathcal{Y}$, respectively, and the training data with $\mathcal{D} = (\boldsymbol{x}_i, y_i)_{i=1}^n$, being n the number of training samples. We use $L(\mathcal{D}, \boldsymbol{w})$ to denote the *loss* incurred by the classifier $f : \mathcal{X} \mapsto \mathcal{Y}$ (parameterized by $\boldsymbol{w}$) on $\mathcal{D}$. We assume that the classification function f is learned by minimizing an objective function $\mathcal{L}(\mathcal{D}, \boldsymbol{w})$ on the training data. Typically, this is an estimate of the generalization error, obtained by the sum of the empirical loss L on $\mathcal{D}$ and a regularization term.

3.1. Attacker's Goal

This aspect is defined in terms of the desired security violation, attack specificity, and error specificity, as detailed below.

Security Violation. The attacker may aim to cause: an *integrity* violation, to evade detection without compromising normal system operation; an *availability* violation, to compromise the normal system functionalities available to legitimate users; or a *privacy* violation, to obtain private information about the system, its users or data by reverse-engineering the learning algorithm.

Attack Specificity. It ranges from *targeted* to *indiscriminate*, respectively, depending on whether the attacker aims to cause misclassification of a specific set of samples (to target a *given* system user or protected service), or of any sample (to target *any* system user or protected service).

Error Specificity. It can be *specific*, if the attacker aims to have a sample misclassified as a specific class; or *generic*, if the attacker aims to have a sample misclassified as any of the classes different from the true class.[2]

3.2. Attacker's Knowledge

The attacker can have different levels of knowledge of the targeted system, including: (*k.i*) the training data $\mathcal{D}$; (*k.ii*) the feature set $\mathcal{X}$; (*k.iii*) the learning algorithm f, along with the objective function $\mathcal{L}$ minimized during training; and, possibly, (*k.iv*) its (trained) parameters/hyper-parameters $\boldsymbol{w}$. The attacker's knowledge can thus be characterized in terms of a space Θ, whose elements encode the components (*k.i*)-(*k.iv*) as $\boldsymbol{\theta} = (\mathcal{D}, \mathcal{X}, f, \boldsymbol{w})$. Depending on the assumptions made on (*k.i*)-(*k.iv*), one can describe different attack scenarios.

Perfect-Knowledge (PK) White-Box Attacks. Here the attacker is assumed to know everything about the targeted system, i.e., $\boldsymbol{\theta}_{\text{PK}} = (\mathcal{D}, \mathcal{X}, f, \boldsymbol{w})$. This setting allows one to perform a worst-case evaluation of the security of learning algorithms,

[2] In [13], the authors defined *targeted* and *indiscriminate* attacks (at test time) depending on whether the attacker aims to cause *specific* or *generic* errors. Here we do not follow their naming convention, as it can cause confusion with the interpretation of *targeted* and *indiscriminate* attack specificity also introduced in previous work [21, 27, 28, 38–40, 57–60].

providing empirical upper bounds on the performance degradation that may be incurred by the system under attack.

Limited-Knowledge (LK) Gray-Box Attacks. One may consider here different settings, depending on the attacker's knowledge about each of the components (*k.i*)-(*k.iv*). Typically, the attacker is assumed to know the feature representation $\mathcal{X}$ and the kind of learning algorithm f (e.g., the fact that the classifier is linear, or it is a neural network with a given architecture, etc.), but neither the training data $\mathcal{D}$ nor the classifier's (trained) parameters $\boldsymbol{w}$. The attacker is however assumed to be able to collect a surrogate data set $\hat{\mathcal{D}}$[3] from a similar source (ideally sampling from the same underlying data distribution), and potentially get feedback from the classifier about its decisions to provide labels for such data. This enables the attacker to estimate the parameters $\hat{\boldsymbol{w}}$ from $\hat{\mathcal{D}}$, by training a *surrogate classifier*. We refer to this case as LK attacks with Surrogate Data (LK-SD), and denote it with $\boldsymbol{\theta}_{\text{LK-SD}} = (\hat{\mathcal{D}}, \mathcal{X}, f, \hat{\boldsymbol{w}})$.

We refer to the setting in which the attacker does not even know the kind of learning algorithm f as LK attacks with Surrogate Learners (LK-SL), and denote it with $\boldsymbol{\theta}_{\text{LK-SL}} = (\hat{\mathcal{D}}, \mathcal{X}, \hat{f}, \hat{\boldsymbol{w}})$. LK-SL attacks also include the case in which the attacker knows the learning algorithm, but optimizing the attack samples against it may be not tractable or too complex. In this case, the attacker can also craft the attacks against a surrogate classifier and test them against the targeted one. This is a common procedure used also to evaluate the *transferability* of attacks between learning algorithms, as firstly shown in [36] and subsequently in [14] for deep networks.

Zero-Knowledge (ZK) Black-Box Attacks. Recent work has also claimed that machine learning can be threatened without any substantial knowledge of the feature space, the learning algorithm and the training data, if the attacker can query the system in a black-box manner and get feedback on the provided labels or confidence scores [14, 62–65]. This point deserves however some clarification. First, the attacker knows (as any other potential user) that the classifier is designed to perform some task (e.g., object recognition in images, malware classification, etc.), and has to clearly have an idea of which potential transformations to apply to cause some feature changes, otherwise neither change can be inflicted to the output of the classification function, nor any useful information can be extracted from it. For example, if one attacks a malware detector based on dynamic analysis by injecting static code that will never be executed, there will be no impact at all on the classifier's decisions. This means that, although the exact feature representation may be not known to the attacker, at least she knows (or has to get to know) which kind of features are used by the system (e.g., features based on static or dynamic analysis in malware detection). Thus, knowledge of the feature representation may be partial, but not completely absent. This is even more evident for deep networks trained on images, where the attacker knows that the input features *are* the image pixels.

Similar considerations hold for knowledge of the training data. If the attacker knows that the classifier is used for a specific task, it is clear the she also knows which kind of data has been used to train it; for example, if a deep network aims to discriminate among classes of animals, then it is clear that it has been trained on images of such animals. Hence, also in this case the attacker effectively has some knowledge of the training data, even if not of the exact training samples.

We thus characterize this setting as $\boldsymbol{\theta}_{\text{ZK}} = (\hat{\mathcal{D}}, \hat{\mathcal{X}}, \hat{f}, \hat{\boldsymbol{w}})$. Even if surrogate learners are not necessarily used here [62–65], as well as in pioneering work on black-box attacks against machine learning [19, 66], one may anyway learn a surrogate classifier (potentially on a different feature representation) and check whether the crafted attack samples *transfer* to the targeted classifier. Feedback from classifier's decisions on carefully-crafted query samples can then be used to refine the surrogate model, as in [14]. Although the problem of learning a surrogate model while minimizing the number of queries can be casted as an *active learning* problem, to our knowledge well-established *active learning* algorithms have not yet been compared against such recently-proposed approaches [14].

3.3. Attacker's Capability

This characteristic depends on the *influence* that the attacker has on the input data, and on application-specific *data manipulation constraints*.

Attack Influence. It can be causative, if the attacker can manipulate both training and test data, or exploratory, if the attacker can only manipulate test data. These scenarios are more commonly known as *poisoning* and *evasion* attacks [21, 24, 27, 29, 31, 36, 38–40, 57].

Data Manipulation Constraints. Another aspect related to the attacker's capability depends on the presence of application-specific constraints on data manipulation, e.g., to evade malware detection, malicious code has to be modified without compromising its intrusive functionality. This may be done against systems based on static code analysis, by injecting instructions or code that will never be executed [10, 36, 37, 46]. These constraints can be generally accounted for in the definition of the optimal attack strategy by assuming that the initial attack samples $\mathcal{D}_c$ can only be modified according to a space of possible modifications $\Phi(\mathcal{D}_c)$. In some cases, this space can also be mapped in terms of constraints on the feature values of the attack samples; e.g., by imposing that feature values corresponding to occurrences of some instructions in static malware detectors can only be incremented [36, 37, 46].

3.4. Attack Strategy

Given the attacker's knowledge $\boldsymbol{\theta} \in \Theta$ and a set of manipulated attack samples $\mathcal{D}'_c \in \Phi(\mathcal{D}_c)$, the attacker's goal can be defined in terms of an objective function $\mathcal{A}(\mathcal{D}'_c, \boldsymbol{\theta}) \in \mathbb{R}$ which measures how effective the attacks $\mathcal{D}'_c$ are. The optimal attack strategy can be thus given as:

$$\mathcal{D}^{\star}_c \in \arg\max_{\mathcal{D}'_c \in \Phi(\mathcal{D}_c)} \mathcal{A}(\mathcal{D}'_c, \boldsymbol{\theta}) \tag{1}$$

We show in Sect. 4 how this high-level formulation encompasses both evasion and poisoning attacks against *supervised*

[3] We use here the *hat* symbol to denote limited knowledge of a given component.

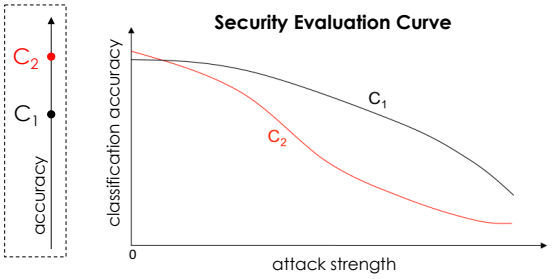

Figure 3: Security evaluation curves of two hypothetical classifiers C_1 and C_2, inspired from the methodology proposed in [39, 40]. Based only on classification accuracy (in the absence of attack), one may prefer C_2 to C_1. Simulating attacks of increasing *maximum strength* (e.g., by increasing the maximum amount of perturbation in input images) may however reveal that more accurate classifiers may be less robust to adversarial input perturbation. After considering the reported security evaluation curves, one may indeed prefer C_1 to C_2.

learning algorithms, despite it has been used also to attack clustering [58–60], and feature selection algorithms [27, 61].

3.5. Security Evaluation Curves

Before delving into the details of specific attacks, we remark that, to provide a thorough security evaluation of machine-learning algorithms, one should assess their performance not only under different assumptions on the attacker's knowledge, but also against increasing *attack strength*, i.e., by increasing the attacker's capability $\Phi(\mathcal{D}_c)$ of manipulating the input data. For example, this can be done by increasing the amount of perturbation used to craft evasion attacks, or the number of poisoning attack points injected into the training data. This is precisely the scope of *security evaluation curves*, which aim to show whether and to which extent the performance of a learning algorithm drops more or less gracefully under attacks of increasing *strength*. This is in fact crucial to enable a fairer comparison among different attack algorithms and defenses, as advocated in [39, 40]. A conceptual example is reported in Fig. 3, while more concrete application examples will be discussed in the remainder of this work.

4. Be Proactive: Simulating Attacks

> *"To know your enemy, you must become your enemy."*
> (Sun Tzu, The Art of War, 500 BC)

We discuss here how to formalize test-time evasion and training-time poisoning attacks in terms of the optimization problem given in Eq. (1), and consistently with the threat model discussed in Sect. 3.[4]

[4]Although we do not thoroughly cover privacy attacks here, we refer the reader to few practical examples of such attacks reported to date, including model inversion attacks aimed to steal machine learning models and earlier hill-climbing attacks against biometric systems used to steal the face and fingerprint templates of their users [28, 62, 67–70].

4.1. Evasion attacks

For evasion attacks, we consider the formulation reported in [6], which extends our previous work [36] from two-class to multiclass classifiers, by introducing *error-generic* and *error-specific* high-confidence evasion attacks. With reference to Eq. (1), the evasion attack samples $\mathcal{D}'_c$ can be optimized one at a time, independently, aiming to maximize the classifier's confidence associated to a wrong class. We will denote with $f_i(\boldsymbol{x})$ the confidence score of the classifier on the sample $\boldsymbol{x}$ for class i. These attacks can be optimized under different levels of attacker's knowledge through the use of surrogate classifiers, so we omit the distinction between $f_i(\boldsymbol{x})$ and $\hat{f}_i(\boldsymbol{x})$ below for notational convenience.

Error-generic Evasion Attacks. In this case, the attacker is interested in misleading classification, regardless of the output class predicted by the classifier. The problem can be thus formulated as:

$$\max_{\boldsymbol{x}'} \quad \mathcal{A}(\boldsymbol{x}', \boldsymbol{\theta}) = \Omega(\boldsymbol{x}') = \max_{l \neq k} f_l(\boldsymbol{x}) - f_k(\boldsymbol{x}) \,, \tag{2}$$

$$\text{s.t.} \quad d(\boldsymbol{x}, \boldsymbol{x}') \leq d_{\max} \,, \;\; \boldsymbol{x}_{\text{lb}} \preceq \boldsymbol{x}' \preceq \boldsymbol{x}_{\text{ub}} \,, \tag{3}$$

where $f_k(\boldsymbol{x})$ denotes the discriminant function associated to the true class k of the source sample $\boldsymbol{x}$, and $\max_{l \neq k} f_l(\boldsymbol{x})$ is the closest competing class (i.e., the one exhibiting the highest value of the discriminant function among the remaining classes). The underlying idea behind this attack formulation, similarly to [4], is to ensure that the attack sample will be no longer classified correctly as a sample of class k, but rather misclassified as a sample of the closest candidate class. The manipulation constraints $\Phi(\mathcal{D}_c)$ are given in terms of: (*i*) a distance constraint $d(\boldsymbol{x}, \boldsymbol{x}') \leq d_{\max}$, which sets a bound on the maximum input perturbation between $\boldsymbol{x}$ (i.e., the input sample) and the corresponding modified adversarial example $\boldsymbol{x}'$; and (*ii*) a box constraint $\boldsymbol{x}_{\text{lb}} \preceq \boldsymbol{x}' \preceq \boldsymbol{x}_{\text{ub}}$ (where $\boldsymbol{u} \preceq \boldsymbol{v}$ means that each element of $\boldsymbol{u}$ has to be not greater than the corresponding element in $\boldsymbol{v}$), which bounds the values of the attack sample $\boldsymbol{x}'$.

For images, the former constraint is used to implement either *dense* or *sparse* evasion attacks [6, 71, 72]. Normally, the ℓ_2 and the ℓ_∞ distances between pixel values are used to cause an indistinguishable image blurring effect (by slightly manipulating all pixels). Conversely, the ℓ_1 distance corresponds to a sparse attack in which only few pixels are significantly manipulated, yielding a salt-and-pepper noise effect on the image [71, 72]. In the image domain, the box constraint can be used to bound each pixel value between 0 and 255, or to ensure manipulation of only a specific region of the image. For example, if some pixels should not be manipulated, one can set the corresponding values of $\boldsymbol{x}_{\text{lb}}$ and $\boldsymbol{x}_{\text{ub}}$ equal to those of $\boldsymbol{x}$. This is of interest to create real-world adversarial examples, as it avoids the manipulation of background pixels which do not belong to the object of interest [6, 16]. Similar constraints have been applied also for evading learning-based malware detectors [36, 37, 46, 71, 72].

Error-specific Evasion Attacks. In the *error-specific* setting, the attacker aims to mislead classification, but she requires the adversarial examples to be misclassified as a specific class. The problem is formulated similarly to error-generic evasion

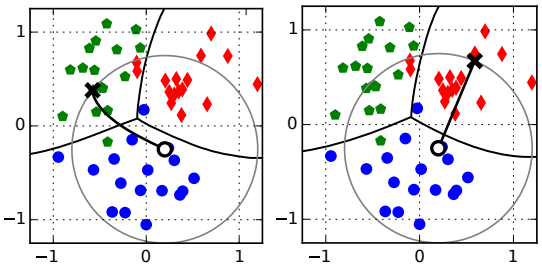

Figure 4: Examples of error-specific (*left*) and error-generic (*right*) evasion [6]. Decision boundaries among the three classes (blue, red and green points) are shown as black lines. In the error-specific case, the initial (blue) sample is shifted towards the green class (selected as target). In the error-generic case, instead, it is shifted towards the red class, as it is the closest class to the initial sample. The gray circle represents the feasible domain, given as an upper bound on the ℓ_2 distance between the initial and the manipulated attack sample.

(Eqs. 2-3), with the only differences that: (*i*) the objective function $\mathcal{A}(\boldsymbol{x}', \boldsymbol{\theta}) = -\Omega(\boldsymbol{x}')$ has opposite sign; and (*ii*) f_k denotes the discriminant function associated to the targeted class, i.e., the class which the adversarial example should be (wrongly) assigned to. The rationale in this case is to maximize the confidence assigned to the wrong target class f_k, while minimizing the probability of correct classification [4, 6].

Attack Algorithm. The two evasion settings are conceptually depicted in Fig. 4. Both can be solved through a straightforward gradient-based attack, for differentiable learning algorithms (including neural networks, SVMs with differentiable kernels, etc.) [6, 36]. Non-differentiable learning algorithms, like decision trees and random forests, can be attacked with more complex strategies [73] or using the same algorithm against a differentiable surrogate learner [72].

4.1.1. Application Example

We report here an excerpt of the results from our recent work [6], where we have constructed adversarial examples aimed to fool the robot-vision system of the iCub humanoid.[5] This system uses a deep network to compute a set of deep features from input images (i.e., by extracting the output of the penultimate layer of the network), and then learns a multiclass classifier on this representation for recognizing 28 different objects, including cups, detergents, hair sprayers, etc. The results for error-specific evasion (averaged on different target classes) are reported in Fig. 5, along with some examples of perturbed input images at different levels. We trained multiclass linear SVMs (SVM), SVMs with the RBF kernel (SVM-RBF), and also a simple defense mechanism against adversarial examples based on rejecting samples that are sufficiently far (in deep space) from known training instances (SVM-adv). This will be discussed more in detail in Sect. 5.2.1 (see also Fig. 9 for a conceptual representation of this defense mechanism). The security evaluation curves in Fig. 5 show how classification accuracy decreases against an increasing ℓ_2 maximum admissible perturbation $d_{\max}$. Notably, the rejection mechanism of SVM-adv is only effective for low input perturbations (at the cost of some additional misclassifications in the absence of attack). For higher perturbation levels, the deep features of the manipulated attacks become indistinguishable to those of the samples of the targeted class, although the input image is still far from resembling a different object. This phenomenon is connected to the instability of the deep representation learned by the underlying deep network. We refer the reader to [6] for further details, and to [74] (and references therein) for the problem of generating adversarial examples in the physical world.

[5] `http://www.icub.org`

4.1.2. Historical Remarks

We conclude this section with some historical remarks on evasion attacks, with the goal of providing a better understanding of the connections with recent work on adversarial examples and the security of deep learning.

Evasion attacks have a long tradition. As mentioned in Sect. 1, back in 2004-2006, work in [19, 20, 32, 75] reported preliminary attempts in evading statistical anti-spam filters and malware detectors with ad-hoc evasion strategies. The very first evasion attacks against linear classifiers were systematized in the same period in [18–20], always considering spam filtering as a running example. The underlying idea was to manipulate the content of spam emails by obfuscating *bad* words and/or adding *good* words. To reduce the number of manipulated words in each spam, and preserve message readability, the idea was to modify first words which were assigned the highest absolute weight values by the linear text classifier. Heuristic countermeasures were also proposed before 2010 [33, 42, 76], based on the intuition of learning linear classifiers with more *uniform* feature weights, to require the attacker to modify more words to get her spam misclassified. To summarize, the vulnerability of linear classifiers to evasion attacks was a known problem even prior to 2010, and simple, heuristic countermeasures were already under development. Meanwhile, Barreno et al. (see [21, 38] and references therein) were providing an initial overview of the vulnerabilities of machine learning from a more general perspective, highlighting the need for *adversarial* machine learning, i.e., to develop learning algorithms that explicitly account for the presence of the attacker [57].

At that time, the idea that nonlinear classifiers could be more robust than linear ones against evasion was also becoming popular. In 2013, Šrndić and Laskov [77] proposed a learning-based PDF malware detector, and attacked it to test its vulnerability to evasion. They reported that:

> *The most aggressive evasion strategy we could conceive was successful for only 0.025% of malicious examples tested against a nonlinear SVM classifier with the RBF kernel [...] we do not have a rigorous mathematical explanation for such a surprising robustness. Our intuition suggests that [...] the space of true features is hidden behind a complex nonlinear transformation which is mathematically hard to invert. [...] the same attack staged against the linear classifier had a 50% success rate; hence, the robustness of the RBF classifier must be rooted in its nonlinear transformation.*

Today we know that this hypothesis about the robustness of nonlinear classifiers is wrong. The fact that a system could be

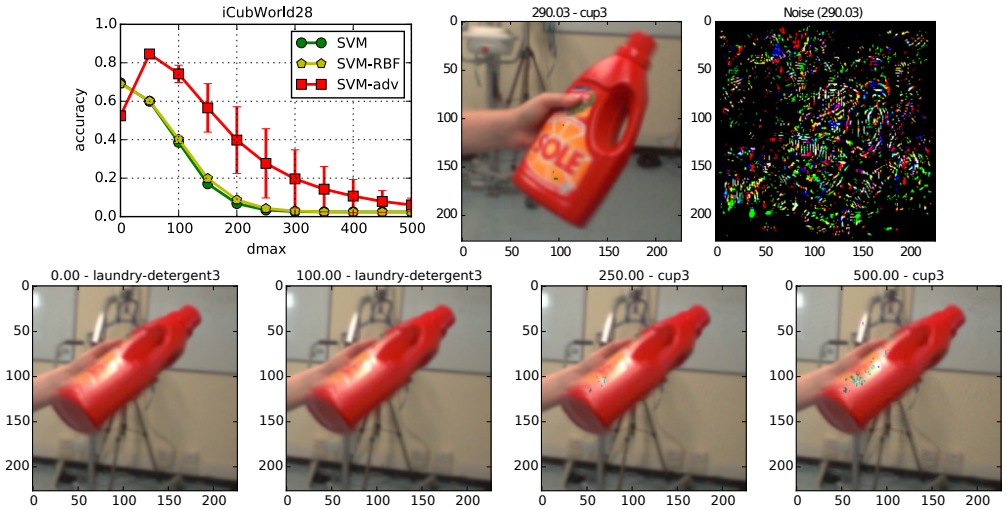

Figure 5: Error-specific evasion results from [6]. *Top row:* Security evaluation curves reporting accuracy of the given classifiers against an increasing ℓ_2 input perturbation. The right-hand side plots depict a laundry detergent misclassified as a cup when applying the minimum input perturbation required for misclassification, along with the corresponding magnified noise mask. *Bottom row:* Images of the laundry detergent perturbed with an increasing level of noise. The manipulations are only barely visible for perturbation values higher than 150-200 (recall however that these values depend on the image size, as the ℓ_2 distance).

more secure against an attack not specifically targeted against it does not provide any further meaningful information about its security to more powerful worst-case attacks. Different systems (and algorithms) should be tested under the same (worst-case) assumptions on the underlying threat model. In particular, it is not difficult to see that the attack developed in that work was somehow crafted to evade linear classifiers, but not sufficiently complex to fool nonlinear ones.

While reading that work, it was thus natural to ask ourselves: "what if the attack is carefully-crafted against nonlinear classifiers, instead? How can we invert such complex nonlinear transformation to understand which features are more relevant to the classification of a sample, and change them?" The answer to this well-posed question was readily available: the gradient of the classification function is exactly what specifies the direction of maximum variation of the function with respect to the input features. Thus, we decided to formulate the evasion of a nonlinear classifier similarly to what we did in [76] for linear classifiers, in terms of an optimization problem that minimizes the discriminant function $f(\boldsymbol{x})$ such that $\boldsymbol{x}$ is misclassified as legitimate with maximum confidence, under a maximum amount of possible changes to its feature vector.

In a subsequent paper [36], we implemented the aforementioned strategy and showed how to evade nonlinear SVMs and neural networks through a straightforward gradient-descent attack algorithm. In the same work, we also reported the first "adversarial examples" on MNIST handwritten digit data against nonlinear learning algorithms. We furthermore showed that, when the attacker does not have perfect knowledge of the targeted classifier, a surrogate classifier can be learned on surrogate training data, and used to craft the attack samples which then *transfer* with high probability to the targeted model. This was also the first experiment showing that adversarial examples can be transferred, at least in a gray-box setting (training the same algorithm on different data). Notably, Šrndić and Laskov [37] subsequently exploited this attack to show that PDF malware detectors based on nonlinear learning algorithms were also vulnerable to evasion, conversely to what they supposed in [77].

More recently, we have also exploited the theoretical findings in [78], which connect regularization and robustness in kernel-based classifiers, to provide a theoretically-sound countermeasure for linear classifiers against evasion attacks [46, 71]. These recent developments have enabled a deeper understanding on how to defend against evasion attacks in spam filtering and malware detection, also clarifying (in a formal manner) the intuitive idea of *uniform* feature weights only heuristically provided in [42, 76]. In particular, we have recently shown how a proper, theoretically-grounded regularization scheme can significantly outperform heuristic approaches in these contexts [46, 71].

Security of Deep Learning. In 2014-2015, Szegedy et al. [1] and subsequent work [2–4] showed that deep networks can be fooled by well-crafted, minimally-perturbed input images at test time, called *adversarial examples*. This instability of deep networks to input perturbations has raised an enormous interest in both the computer vision and security communities which, since then, have started proposing novel security assessment methodologies, attacks and countermeasures to mitigate this threat, almost regardless of previous work done in the area of adversarial machine learning and, in particular, related to evasion attacks. Notably, Papernot et al. [13] proposed an attack framework specifically aimed to assess the security of deep networks, and a defense mechanism (referred to as *distillation*) based on masking the gradient of deep networks to make gradient-based attacks against them ineffective [5]. The same authors subsequently discovered that distillation was vulnerable to attacks crafted against surrogate classifiers with smoother decision functions [14], essentially leveraging the idea behind limited-knowledge evasion attacks we first discussed in [36].

Misconceptions on Evasion Attacks. The main misconception that is worth highlighting here is that *adversarial examples should be minimally perturbed*. The motivation of this miscon-

Figure 6: Timeline of evasion attacks (i.e., adversarial examples) in adversarial machine learning, compared to work on the security of deep networks. Related work is highlighted with markers of the same color, as reported in the legend.

ception is easy to explain. The notion of adversarial examples was initially introduced to analyze the instability of deep networks [1], i.e., to analyze the sensitivity of deep network to minimal changes of the inputs; the goal of the initial work on

adversarial examples was not to to perform a detailed security assessment of a machine learning algorithm using security evaluation curves (Fig. 5). Normally, as already discussed in this paper and also in our previous work [36, 39], for the purpose of thoroughly assessing the security of a learning algorithm under attack, given a feasible space of modifications to the input data, it is more reasonable to assume that the attacker will aim to maximize the classifier's confidence on the desired output class, rather than only minimally perturbing the attack samples (cf. Eqs. 2-3). Minimally-perturbed adversarial examples are adequate to analyze the sensitivity of a machine learning algorithm, but not to thoroughly assess the robustness of a learning algorithm against attacks performed by a rationale attacker, who can perturbate the input to a much larger extent than the minimum required to evade detection. The use of the security evaluation curves described above is thus necessary to assess the security of a learning algorithm against an attacker who can perturb inputs to a much larger extent. In fact, by increasing the attack strength (i.e., the feasible space of modifications to the input data), one can draw a complete security evaluation curve (reporting the evasion rate against an increasing amount of input data perturbation), thus providing a more thorough understanding of system security. This is witnessed by the work by Carlini and Wagner [15, 79], who exploited a similar idea to show that several recent defenses proposed against minimally-perturbed adversarial examples are vulnerable to high-confidence ones, using a stronger attack similar to those proposed in our earlier work, and discussed in Sect. 4.1 [36, 39]. Even in the domain of malware detection, adversarial examples seem to be a novel threat [10], while the vulnerability of learning-based malware detectors to evasion is clearly a consolidated issue [36, 37, 46]. Another interesting avenue to provide reliable guarantees on the security of neural networks is *formal verification*, which however has only been considered for simple network architectures [80]. Other evaluation methodologies leverage ideas from the field of software testing [81].

Timeline of Evasion Attacks. To summarize, while the security of deep networks has received considerable attention from different research communities only recently, it is worth remarking that several related problems and solutions had been already considered prior to 2014 in the field of adversarial machine learning. High-confidence evasion attacks and surrogate models are just two examples of similar findings in both areas of research. We compactly and conceptually highlight these connections in the timeline reported in Fig. 6.[6]

4.2. Poisoning Attacks

As done for evasion attacks, we discuss here error-generic and error-specific poisoning attacks in a PK white-box setting, given that the extension to gray-box and black-box settings is trivial through the use of surrogate learners [31].

[6] An online version of the timeline is also available at: `https://sec-ml.pluribus-one.it`, along with a web application that allows one to generate adversarial examples and evaluate if they are able to evade detection (evasion attacks).

Error-Generic Poisoning Attacks. In this case, the attacker aims to cause a *denial of service*, by inducing as many misclassifications as possible (regardless of the classes in which they occur). Poisoning attacks are generally formulated as bilevel optimization problems, in which the outer optimization maximizes the attacker's objective $\mathcal{A}$ (typically, a loss function L computed on untainted data), while the inner optimization amounts to learning the classifier on the poisoned training data [24, 27, 29]. This can be made explicit by rewriting Eq. (1) as:

$$\mathcal{D}_c^\star \in \arg\max_{\mathcal{D}'_c \in \Phi(\mathcal{D}_c)} \quad \mathcal{A}(\mathcal{D}'_c, \boldsymbol{\theta}) = L(\mathcal{D}_{\text{val}}, \boldsymbol{w}^\star)\,, \tag{4}$$

$$\text{s.t.} \quad \boldsymbol{w}^\star \in \arg\min_{\boldsymbol{w}' \in \mathcal{W}} \mathcal{L}(\mathcal{D}_{\text{tr}} \cup \mathcal{D}'_c, \boldsymbol{w}')\,, \tag{5}$$

where $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{val}}$ are two data sets available to the attacker. The former, along with the poisoning attack samples $\mathcal{D}'_c$, is used to train the learner on poisoned data, while the latter is used to evaluate its performance on untainted data, through the loss function $L(\mathcal{D}_{\text{val}}, \boldsymbol{w}^\star)$. Notably, the objective function implicitly depends on $\mathcal{D}'_c$ through the parameters $\boldsymbol{w}^\star$ of the poisoned classifier.

Error-Specific Poisoning Attacks. In this setting the attacker aims to cause specific misclassifications. While the problem remains that given by Eqs. (4)-(5), the objective is redefined as $\mathcal{A}(\mathcal{D}'_c, \boldsymbol{\theta}) = -L(\mathcal{D}'_{\text{val}}, \boldsymbol{w}^\star)$. The set $\mathcal{D}'_{\text{val}}$ contains the same samples as $\mathcal{D}_{\text{val}}$, but their labels are chosen by the attacker according to the desired misclassifications. The objective L is then taken with opposite sign as the attacker effectively aims to *minimize* the loss on her desired labels [31].

Attack Algorithm. A common trick used to solve the given bilevel optimization problems is to replace the inner optimization by its equilibrium conditions [24, 27, 29, 31]. This enables gradient computation in closed form and, thus, similarly to the evasion case, the derivation of gradient-based attacks (although gradient-based poisoning is much more computationally demanding, as it requires retraining the classifier iteratively on the modified attack samples). In the case of deep networks, this approach is not practical due to computational complexity and instability of the closed-form gradients. To tackle this issue, we have recently proposed a more efficient technique, named *back-gradient poisoning*. It relies on automatic differentiation and on reversing the learning procedure to compute the gradient of interest (see [31] for further details).

4.2.1. Application Example

We report here an exemplary poisoning attack against a multiclass softmax classifier (logistic regression) trained on MNIST handwritten digits belonging to class 0, 4, and 9. We consider error-generic poisoning, using 200 (clean) training samples and 2000 validation and test samples. Results of back-gradient poisoning compared to randomly-injected training points with wrong class labels (random label flips) are reported in Fig. 7, along with some *adversarial training examples* generated by our back-gradient poisoning algorithm.

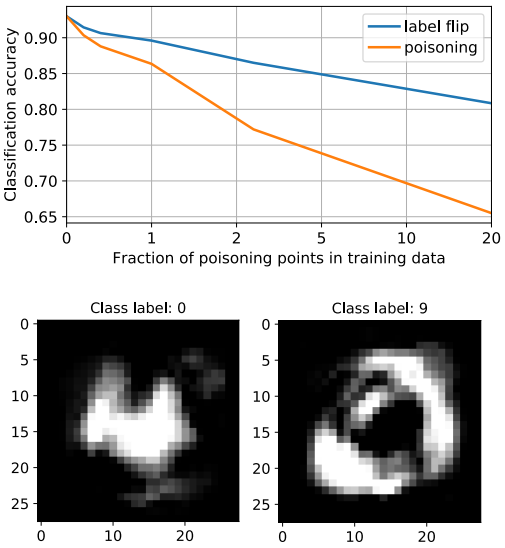

Figure 7: *Top row:* Security evaluation curve of a softmax classifier trained on the MNIST digits 0, 4, and 9, against back-gradient poisoning and random label flips (baseline comparison). *Bottom row:* Examples of adversarial training digits generated by back-gradient poisoning.

4.2.2. Historical Remarks

To our knowledge, the earliest poisoning attacks date back to 2006-2010 [21–23, 25, 82]. Newsome et al. [82] devised an attack to mislead signature generation for malware detection; Nelson et al. [22] showed that spam filters can be compromised to misclassify legitimate email as spam, by learning spam emails containing *good* words during training; and Rubinstein et al. [23] showed how to poison an anomaly detector trained on network traffic through injection of *chaff* traffic. In the meanwhile, exemplary attacks against learning-based centroid anomaly detectors where also demonstrated [21, 25, 26]. Using a similar formalization, we have also recently showed poisoning attacks against biometric systems [41]. This background paved the way to subsequent work that formalized poisoning attacks against more complex learning algorithms (including SVMs, ridge regression, and LASSO) as bilevel optimization problems [24, 27, 29]. Recently, preliminary attempts towards poisoning deep networks have also been reported, showing the first *adversarial training examples* against deep learners [30, 31].

It is worth finally remarking that poisoning attacks against machine learning should not be considered an academic exercise in vitro. Microsoft *Tay*, a chatbot designed to talk to youngsters in Twitter, was shut down after only 16 hours, as it started raising racist and offensive comments after being poisoned.[7] Its artificial intelligence was designed to mimic the behavior of humans, but not to recognize potential misleading behaviors. Kaspersky Lab, a leading antivirus company, has been accused of poisoning competing antivirus products through the injection of false positive examples into VirusTotal,[8] although it is worth saying that they denied any wrongdoing, and blamed for spreading false rumors. Another avenue for poisoning arises from the fact that shared, big and open data sets are commonly used to train machine-learning algorithms. The case of ImageNet for object recognition is paradigmatic. In fact, people typically reuse these large-scale deep networks as feature extractors inside their pattern recognition tools. Imagine what may happen if someone could poison these data "reservoirs": many data-driven products and services could experience security and privacy issues, economic losses, with legal and ethical implications.

[7] `http://wired.com/2017/02/keep-ai-turning-racist-monster`
[8] `http://virustotal.com`

5. Protect Yourself: Security Measures for Learning Algorithms

> *"What is the rule? The rule is protect yourself at all times."*
>
> (from the movie *Million dollar baby*, 2004)

In this section we discuss the *third golden rule* of the security-by-design cycle for pattern classifiers, i.e., how to react to *past* attacks and prevent *future* ones. We categorize the corresponding defenses as depicted in Fig. 8.

5.1. Reactive Defenses

Reactive defenses aim to counter *past* attacks. In some applications, reactive strategies may be even more convenient and effective than pure proactive approaches aimed to solely mitigate the risk of potential future attacks [28, 40, 83]. Reactive approaches include: (*i*) timely detection of novel attacks, (*ii*) frequent classifier retraining, and (*iii*) verification of consistency of classifier decisions against training data and ground-truth labels [40, 49]. In practice, to timely identify and block novel security threats, one can leverage collaborative approaches and honeypots, i.e., online services purposely vulnerable with the specific goal of collecting novel spam and malware samples. To correctly detect recently-reported attacks, the classifier should be frequently retrained on newly-collected data (including them), and novel features and attack detectors may also be considered (see, e.g., the spam arms race discussed in Sect. 2). This procedure should also be automated to some extent to act more readily when necessary; e.g., using automatic *drift* detection techniques [11, 40, 84]. The correctness of classifier decisions should finally be verified by expert domains. This raises the issue of how to involve *humans in the loop* in a more coordinated manner, to supervise and verify the correct functionality of learning systems.

5.2. Proactive Defenses

Proactive defenses aim to prevent *future* attacks. The main ones proposed thus far can be categorized according to the paradigms of *security by design* and *security by obscurity*, as discussed in the following.

5.2.1. Security-by-Design Defenses against White-box Attacks

The paradigm of security by design advocates that a system should be designed from the ground up to be secure. Based on this idea, several learning algorithms have been adapted to explicitly take into account different kinds of adversarial data manipulation. These defenses are designed in a *white-box* setting

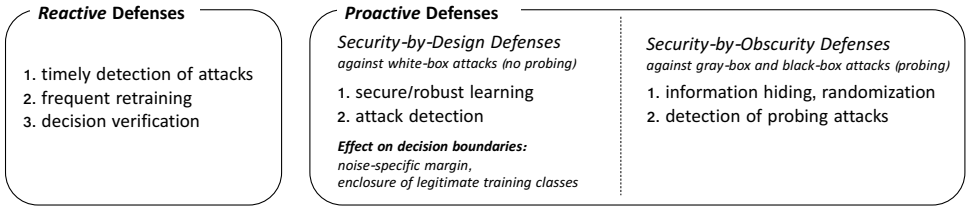

Figure 8: Schematic categorization of the defense techniques discussed in Sect. 5.

in which the attacker is assumed to have perfect knowledge of the attacked system. There is thus no need to probe the targeted classifier to improve knowledge about its behavior (as instead done in gray-box and black-box attacks).

Countering Evasion Attacks. In 2004, Dalvi et al. [18] proposed the first adversary-aware classifier against evasion attacks, based on iteratively retraining the classifier on the simulated attacks. This is not very different from the idea of *adversarial training* that has been recently used in deep networks to counter adversarial examples [1, 2], or to harden decision trees and random forests [73]. These defenses are however heuristic, with no formal guarantees on convergence and robustness properties. More theoretically-sound approaches relying on *game theory* have been proposed to overcome these limitations. Zero-sum games have been formulated to learn *invariant* transformations like feature insertion, deletion and rescaling [33–35]. Then, more rigorous approaches have introduced Nash and Stackelberg games for secure learning, deriving formal conditions for existence and uniqueness of the game equilibrium, under the assumption that each player knows everything about the opponents and the game [44, 85]. Randomized players [45] and uncertainty on the players' strategies [86] have also been considered to simulate less pessimistic scenarios. Despite these approaches seem promising, understanding the extent to which the resulting attack strategies are representative of practical scenarios remains an open issue [87, 88]. Adversarial learning is not a (board) game with well-defined rules and, thus, the objective functions of real-world attackers may not even correspond to those hypothesized in the aforementioned games. It may be thus interesting to verify, reactively, whether real-world attackers behave as hypothesized, and exploit feedback from the observed attacks to improve the definition of the attack strategy. Another relevant problem of these approaches is their scalability to large datasets and high-dimensional feature spaces, as it may be too computationally costly to generate a sufficient number of attack samples to correctly represent their distribution, i.e., to effectively tackle the curse of dimensionality.

A more efficient approach, similar to game-theoretical ones, relies on *robust optimization*, in which adversarial data manipulation can be seen as a particular kind of noise. In particular, Xu et al. [78] have shown that different regularizers amount to hypothesizing different kinds of bounded worst-case noise on the input data, at least for kernel-based classifiers. This has effectively established an equivalence between regularized learning problems and robust optimization, which has in turn enabled approximating computationally-demanding secure learning models (e.g., game-theoretical ones) with more efficient ones based on regularizing the objective function in a specific manner [46, 71, 72], also in structured learning [89]. Hybrid approaches based on regularizing gradients through simulation of the corresponding attacks have also been recently proposed to improve the security of deep networks to evasion attacks [90, 91].

Another line of defenses against evasion attacks is based on detecting and rejecting samples which are sufficiently far from the training data in feature space (similarly to the defense discussed in Sect. 4.1.1) [6, 7, 11, 92, 93]. These samples are usually referred to as *blind-spot* evasion points, as they appear in regions of the feature space scarcely populated by training data. These regions can be assigned to any class during classifier training without any substantial increase in the training loss. In practice, this is a simple consequence of the *stationarity* assumption underlying many machine-learning algorithms (according to which training and test data come from the same distribution), and such rejection-based defenses simply aim to overcome this issue.

Finally, we point out that *classifier ensembles* have been also exploited to improve security against evasion attempts (e.g., by implementing rejection-based mechanisms) [42, 53, 76, 92] and even against poisoning attacks [94]. They may however worsen security if the base classifiers are not properly *combined* [53, 92].

Effect on Decision Boundaries. We aim to discuss here how the proposed defenses substantially *change* the way classifiers learn their decision boundaries. Notably, defenses involving retraining on the attack samples and rejection mechanisms achieve security against *evasion* by essentially countering blind-spot attacks. One potential effect of this assumption is that the resulting decision functions may tend to *enclose* the (stationary) training classes more tightly. This in turn may require one to trade-off between the security against potential attacks and the number of misclassified (stationary) samples at test time, as empirically shown in Sect. 4.1.1, and conceptually depicted in Fig. 9 [6]. The other relevant effect, especially induced by regularization methods inspired from robust optimization, is to provide a *noise-specific margin* between classes, as conceptually represented in Fig. 10 [71, 72]. These are the two main effects induced by the aforementioned secure learning approaches in feature space.

It is finally worth remarking that, by using a *secure learning* algorithm, one can counter blind-spot evasion samples, but definitely not adversarial examples whose feature vectors become *indistinguishable* from those of training samples belonging to different classes. In this case, indeed, any learning algorithm

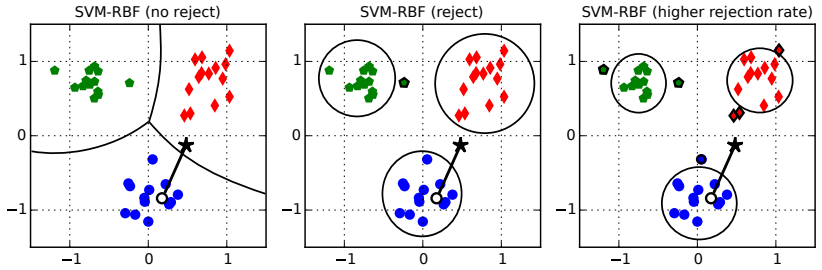

Figure 9: Effect of *class-enclosing* defenses against blind-spot adversarial examples on multiclass SVMs with RBF kernels, adapted from [6]. Rejected samples are highlighted with black contours. The adversarial example (black star) is misclassified only by the standard SVM (left plot), while SVM with rejection correctly identifies it as an adversarial example (middle plot). Rejection thresholds can be modified to increase classifier security by tightening class enclosure (right plot), at the expense of misclassifying more legitimate samples.

would not be able to tell such samples apart [95]. The security properties of learning algorithms should be thus considered independently from those exhibited by the chosen feature representation. Security of features should be considered as an additional, important requirement; features should not only be discriminant, but also *robust* to manipulation, to avoid straightforward classifier evasion by mimicking the feature values exhibited by legitimate samples. In the case of deep convolutional networks, most of the problems arise from the fact that the learned mapping from input to deep space (i.e., the feature representation) violates the smoothness assumption of learning algorithms: samples that are close in input space may be very far in deep space. In fact, as also reported in Sect. 4.1.1, adversarial examples in *deep space* become indistinguishable from training samples of other classes for sufficiently-high adversarial input perturbations [6]. Therefore, this vulnerability can only be patched by retraining or re-engineering the deeper layers of the network (and not only the last ones) [1, 6].

Countering Poisoning Attacks. While most work focused on countering evasion attacks at test time, some white-box defenses have also been proposed against poisoning attacks [22, 23, 94, 96–99]. To compromise a learning algorithm during training, an attack has to be exhibit different characteristics from those shown by the rest of the training data (otherwise it would have no impact at all) [94]. Poisoning attacks can be thus regarded as *outliers*, and countered using data sanitization (i.e., attack detection and removal) [94, 97, 99], and robust learning (i.e., learning algorithms based on *robust statistics* that are intrinsically less sensitive to outlying training samples) [23, 98].

5.2.2. Security-by-Obscurity Defenses against Black-box Attacks

These proactive defenses, also known as *disinformation* techniques in [21, 38, 57], follow the paradigm of *security by obscurity*, i.e., they hide information to the attacker to improve security. These defenses aim to counter gray-box and black-box attacks in which probing mechanisms are used to improve surrogate models or refine evasion attempts by querying the targeted classifier.

Some examples include [49]: (*i*) randomizing collection of training data (collect at different timings, and locations); (*ii*) using difficult to reverse-engineer classifiers (e.g., classifier ensembles); (*iii*) denying access to the actual classifier or training data; and (*iv*) randomizing the classifier's output to give imperfect feedback to the attacker. The latter approach has been firstly proposed in 2008 [43] as an effective way to hide information about the classification function to the attacker, with recent follow-ups in [7, 45] to counter adversarial examples. However, it is still an open issue to understand whether and to which extent randomization may be used to make it harder for the attacker to learn a proper surrogate model, and to implement privacy-preserving mechanisms [100] against model inversion and hill-climbing attacks [28, 62, 67–70].

Notably, security-by-obscurity defenses may not always be helpful. Gradient masking has been proposed to hide the gradient direction used to craft adversarial examples [5, 8], but it has been shown that it can be easily circumvented with surrogate learners [14, 15, 36], exploiting the same principle behind attacking non-differentiable classifiers (discussed in Sect. 4.1) [72].

6. Conclusions and Future Work

In this paper, we have presented a thorough overview of work related to the security of machine learning, pattern recognition, and deep neural networks, with the goal of providing a clearer historical picture along with useful guidelines on how to assess and improve their security against adversarial attacks.

We conclude this work by discussing some future research paths arising from the fact that machine learning has been originally developed for *closed-world* problems where the possible "states of nature" and "actions" that a rationale agent can implement are perfectly known. Using the words of a famous speech by Donald Rumsfeld, one could argue that machine learning can deal with *known unknowns*.[9] Unfortunately, adversarial machine learning often deals with *unknown unknowns*. When learning systems are deployed in adversarial environments in the *open world*, they can misclassify (with high-confidence) never-before-seen inputs that are largely different from known training data. We know that *unknown unknowns* are the real threat in many security problems (e.g., zero-day attacks in computer security). Although they can be mitigated using the proactive approach described in this work, they remain a primary

[9] `http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636`

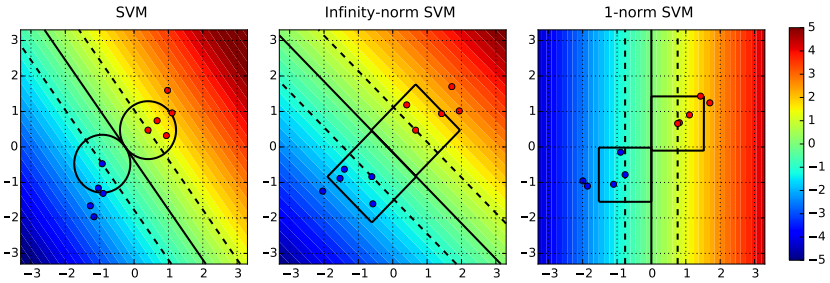

Figure 10: Decision functions for linear SVMs with ℓ_2, ℓ_∞ and ℓ_1 regularization on the feature weights [71, 72]. The feasible domain of adversarial modifications (characterizing the equivalent robust optimization problem) is shown for some training points, respectively with ℓ_2, ℓ_1 and ℓ_∞ balls. Note how the shape of these balls influences the orientation of the decision boundaries, i.e., how different regularizers optimally counter specific kinds of adversarial noise.

open issue for adversarial machine learning, as modeling attacks relies on *known unknowns*, while *unknown unknowns* are unpredictable.

We are firmly convinced that new research paths should be explored to address this fundamental issue, complementary to formal verification and certified defenses [80, 99]. Machine learning algorithms should be able to detect *unknown unknowns* using robust methods for anomaly or novelty detection, potentially asking for human intervention when required. The development of practical methods for explaining, visualizing and interpreting the operation of machine-learning systems could also help system designers to investigate the behavior of such systems on cases that are not statistically represented by the training data, and decide whether to trust their decisions on such *unknown unknowns* or not. These future research paths lie at the intersection of the field of adversarial machine learning and the emerging fields of robust artificial intelligence and interpretability of machine learning [101, 102], and we believe that these directions will help our society to get a more conscious understanding of the potential and limits of modern data-driven AI and machine-learning technologies.

Acknowledgments

We are grateful to Ambra Demontis and Marco Melis for providing the experimental results on evasion and poisoning attacks.

References

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: ICLR, 2014.

[2] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: ICLR, 2015.

[3] A. M. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images., in: IEEE CVPR, 2015, pp. 427–436.

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: IEEE CVPR, 2016, pp. 2574–2582.

[5] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in: IEEE Symp. Security & Privacy (SP), 2016, pp. 582–597.

[6] M. Melis, A. Demontis, B. Biggio, G. Brown, G. Fumera, F. Roli, Is deep learning safe for robot vision? Adversarial examples against the iCub humanoid, in: ICCV Workshop ViPAR, 2017.

[7] D. Meng, H. Chen, MagNet: a two-pronged defense against adversarial examples, in: 24th ACM Conf. Computer and Comm. Sec. (CCS), 2017.

[8] J. Lu, T. Issaranon, D. Forsyth, Safetynet: Detecting and rejecting adversarial examples robustly, in: IEEE ICCV, 2017.

[9] X. Li, F. Li, Adversarial examples detection in deep networks with convolutional filter statistics, in: IEEE ICCV, 2017.

[10] K. Grosse, N. Papernot, P. Manoharan, M. Backes, P. D. McDaniel, Adversarial examples for malware detection, in: ESORICS (2), Vol. 10493 of LNCS, Springer, 2017, pp. 62–79.

[11] R. Jordaney, K. Sharad, S. K. Dash, Z. Wang, D. Papini, I. Nouretdinov, L. Cavallaro, Transcend: Detecting concept drift in malware classification models, USENIX Sec. Symp., USENIX Assoc., 2017, pp 625?642.

[12] P. McDaniel, N. Papernot, Z. B. Celik, Machine learning in adversarial settings, IEEE Security & Privacy 14 (3) (2016) 68–72.

[13] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: 1st IEEE European Symp. Security and Privacy, 2016, pp. 372–387.

[14] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, A. Swami, Practical black-box attacks against machine learning, in: ASIA CCS '17, ACM, 2017, pp. 506–519.

[15] N. Carlini, D. A. Wagner, Towards evaluating the robustness of neural networks, in: IEEE Symp. Security & Privacy (SP), 2017, pp. 39–57.

[16] M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter, Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, in: Conf. Computer and Comm. Security (CCS), ACM, 2016, pp. 1528–1540.

[17] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, A. Yuille, Adversarial examples for semantic segmentation and object detection, in: IEEE ICCV, 2017.

[18] N. Dalvi, P. Domingos, Mausam, S. Sanghai, D. Verma, Adversarial classification, in: Int'l Conf. Knowl. Disc. and Data Mining, 2004, pp. 99–108.

[19] D. Lowd, C. Meek, Adversarial learning, in: Int'l Conf. Knowl. Disc. and Data Mining, ACM Press, Chicago, IL, USA, 2005, pp. 641–647.

[20] D. Lowd, C. Meek, Good word attacks on statistical spam filters, in: 2nd Conf. Email and Anti-Spam (CEAS), Mountain View, CA, USA, 2005.

[21] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, J. D. Tygar, Can machine learning be secure?, in: ASIA CCS '06, ACM, 2006, pp. 16–25.

[22] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, K. Xia, Exploiting machine learning to subvert your spam filter, in: LEET '08, USENIX Assoc., 2008, pp. 1–9.

[23] B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, J. D. Tygar, Antidote: understanding and defending against poisoning of anomaly detectors, in: IMC '09, ACM, 2009, pp. 1–14.

[24] B. Biggio, B. Nelson, P. Laskov, Poisoning attacks against support vector machines, in: 29th ICML, 2012, pp. 1807–1814.

[25] M. Kloft, P. Laskov, Online anomaly detection under adversarial impact, in: 13th AISTATS, 2010, pp. 405–412.

[26] M. Kloft, P. Laskov, Security analysis of online centroid anomaly detection, JMLR 13 (2012) 3647–3690.

[27] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, F. Roli, Is feature selection secure against training data poisoning?, in: 32nd ICML, Vol. 37,

2015, pp. 1689–1698.
[28] B. Biggio, I. Corona, B. Nelson, B. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, F. Roli, Security evaluation of support vector machines in adversarial environments, in: Y. Ma, G. Guo (Eds.), Support Vector Machines Applications, Springer Int'l Publishing, Cham, 2014, pp. 105–153.
[29] S. Mei, X. Zhu, Using machine teaching to identify optimal training-set attacks on machine learners, in: 29th AAAI, 2015.
[30] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: ICML, 2017.
[31] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, F. Roli, Towards poisoning of deep learning algorithms with back-gradient optimization, in: AISec '17, ACM, 2018, pp. 27–38.
[32] G. L. Wittel, S. F. Wu, On attacking statistical spam filters, in: 1st Conf. Email and Anti-Spam (CEAS), 2004.
[33] A. Globerson, S. T. Roweis, Nightmare at test time: robust learning by feature deletion, in: 23rd ICML, Vol. 148, ACM, 2006, pp. 353–360.
[34] C. H. Teo, A. Globerson, S. Roweis, A. Smola, Convex learning with invariances, in: NIPS 20, MIT Press, 2008, pp. 1489–1496.
[35] O. Dekel, O. Shamir, L. Xiao, Learning to classify with missing and corrupted features, Machine Learning 81 (2010) 149–178.
[36] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: ECML PKDD, Part III, Vol. 8190 of LNCS, Springer, 2013, pp. 387–402.
[37] N. Šrndic, P. Laskov, Practical evasion of a learning-based classifier: A case study, in: IEEE Symp. Security and Privacy, SP '14, 2014, pp. 197–211.
[38] M. Barreno, B. Nelson, A. Joseph, J. Tygar, The security of machine learning, Machine Learning 81 (2010) 121–148.
[39] B. Biggio, G. Fumera, F. Roli, Security evaluation of pattern classifiers under attack, IEEE Trans. Knowl. and Data Eng. 26 (4) (2014) 984–996.
[40] B. Biggio, G. Fumera, F. Roli, Pattern recognition systems under attack: Design issues and research challenges, IJPRAI 28 (7) (2014) 1460002.
[41] B. Biggio, G. Fumera, P. Russu, L. Didaci, F. Roli, Adversarial biometric recognition : A review on biometric system security from the adversarial machine-learning perspective, IEEE Signal Proc. Mag., 32 (5) (2015) 31–41.
[42] A. Kolcz, C. H. Teo, Feature weighting for improved classifier robustness, in: 6th Conf. Email and Anti-Spam (CEAS), 2009.
[43] B. Biggio, G. Fumera, F. Roli, Adversarial pattern classification using multiple classifiers and randomisation, in: SSPR 2008, Vol. 5342 of LNCS, Springer, 2008, pp. 500–509.
[44] M. Brückner, C. Kanzow, T. Scheffer, Static prediction games for adversarial learning problems, JMLR 13 (2012) 2617–2654.
[45] S. Rota Bulò, B. Biggio, I. Pillai, M. Pelillo, F. Roli, Randomized prediction games for adversarial machine learning, IEEE Trans. Neural Networks and Learning Systems 28 (11) (2017) 2466–2478.
[46] A. Demontis, M. Melis, B. Biggio, D. Maiorca, D. Arp, K. Rieck, I. Corona, G. Giacinto, F. Roli, Yes, machine learning can be more secure! A case study on android malware detection, IEEE Trans. Dep. and Secure Comp.
[47] P. Laskov, R. Lippmann (Eds.), NIPS Workshop on Machine Learning in Adversarial Environments for Computer Security, 2007.
[48] P. Laskov, R. Lippmann, Machine learning in adversarial environments, Machine Learning 81 (2010) 115–119.
[49] A. D. Joseph, P. Laskov, F. Roli, J. D. Tygar, B. Nelson, Machine Learning Methods for Computer Security (Dagstuhl Perspectives Workshop 12371), Dagstuhl Manifestos 3 (1) (2013) 1–30.
[50] B. M. Thuraisingham, B. Biggio, D. M. Freeman, B. Miller, A. Sinha (Eds.), AISec '17: 10th Workshop on AI and Security, ACM, 2017.
[51] A. D. Joseph, B. Nelson, B. I. P. Rubinstein, J. Tygar, Adversarial Machine Learning, Cambridge University Press, 2018.
[52] X. Han, N. Kheir, D. Balzarotti, Phisheye: Live monitoring of sandboxed phishing kits, in: ACM CCS, 2016, pp. 1402–1413.
[53] I. Corona, B. Biggio, M. Contini, L. Piras, R. Corda, M. Mereu, G. Mureddu, D. Ariu, F. Roli, Deltaphish: Detecting phishing webpages in compromised websites, in: ESORICS, Vol. 10492 of LNCS, Springer, 2017, pp. 370–388.
[54] B. Biggio, G. Fumera, I. Pillai, F. Roli, A survey and experimental evaluation of image spam filtering techniques, PRL 32 (10) (2011) 1436 – 1446.
[55] A. Attar, R. M. Rad, R. E. Atani, A survey of image spamming and filtering techniques, Artif. Intell. Rev. 40 (1) (2013) 71–105.
[56] G. Fumera, I. Pillai, F. Roli, Spam filtering based on the analysis of text information embedded into images, JMLR 7 (2006) 2699–2720.
[57] L. Huang, A. D. Joseph, B. Nelson, B. Rubinstein, J. D. Tygar, Adversarial machine learning, in: 4th AISec, Chicago, IL, USA, 2011, pp. 43–57.
[58] B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, F. Roli, Is data clustering in adversarial settings secure?, in: AISec '13, ACM, 2013, pp. 87–98.
[59] B. Biggio, K. Rieck, D. Ariu, C. Wressnegger, I. Corona, G. Giacinto, F. Roli, Poisoning behavioral malware clustering, in: AISec '14, ACM, 2014, pp. 27–36.
[60] B. Biggio, S. R. Bulò, I. Pillai, M. Mura, E. Z. Mequanint, M. Pelillo, F. Roli, Poisoning complete-linkage hierarchical clustering, in: SSPR, Vol. 8621 of LNCS, Springer, 2014, pp. 42–52.
[61] F. Zhang, P. Chan, B. Biggio, D. Yeung, F. Roli, Adversarial feature selection against evasion attacks, IEEE Trans. Cyb. 46 (3) (2016) 766–777.
[62] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, T. Ristenpart, Stealing machine learning models via prediction APIs, in: USENIX Sec. Symp., USENIX Assoc., 2016, pp. 601–618.
[63] W. Xu, Y. Qi, D. Evans, Automatically evading classifiers, in: Annual Network & Distr. Sys. Sec. Symp. (NDSS), The Internet Society, 2016.
[64] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, C.-J. Hsieh, Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: AISec '17, ACM, 2017, pp. 15–26.
[65] H. Dang, Y. Huang, E. Chang, Evading classifiers by morphing in the dark, in: ACM CCS '17, ACM, 2017, pp. 119–133.
[66] B. Nelson, B. I. Rubinstein, L. Huang, A. D. Joseph, S. J. Lee, S. Rao, J. D. Tygar, Query strategies for evading convex-inducing classifiers, JMLR 13 (2012) 1293–1332.
[67] M. Fredrikson, S. Jha, T. Ristenpart, Model inversion attacks that exploit confidence information and basic countermeasures, in: ACM CCS '15, ACM, 2015, pp. 1322–1333.
[68] A. Adler, Vulnerabilities in biometric encryption systems, in: T. Kanade, A. K. Jain, N. K. Ratha (Eds.), 5th Int'l Conf. Audio- and Video-Based Biometric Person Auth., Vol. 3546 of LNCS, Springer, 2005, pp. 1100–1109.
[69] J. Galbally, C. McCool, J. Fierrez, S. Marcel, J. Ortega-Garcia, On the vulnerability of face verification systems to hill-climbing attacks, Patt. Rec. 43 (3) (2010) 1027–1038.
[70] M. Martinez-Diaz, J. Fierrez, J. Galbally, J. Ortega-Garcia, An evaluation of indirect attacks and countermeasures in fingerprint verification systems, Patt. Rec. Lett. 32 (12) (2011) 1643 – 1651.
[71] A. Demontis, P. Russu, B. Biggio, G. Fumera, F. Roli, On security and sparsity of linear classifiers for adversarial settings, in: SSPR, Vol. 10029 of LNCS, Springer, 2016, pp. 322–332.
[72] P. Russu, A. Demontis, B. Biggio, G. Fumera, F. Roli, Secure kernel machines against evasion attacks, in: AISec '16, ACM, 2016, pp. 59–69.
[73] A. Kantchelian, J. D. Tygar, A. D. Joseph, Evasion and hardening of tree ensemble classifiers, in: ICML, Vol. 48 JMLR W&CP, 2016, pp. 2387–2396.
[74] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, Synthesizing robust adversarial examples, in: ICLR, 2018.
[75] P. Fogla, M. Sharif, R. Perdisci, O. Kolesnikov, W. Lee, Polymorphic blending attacks, in: USENIX Sec. Symp., USENIX Assoc., 2006, pp. 241–256.
[76] B. Biggio, G. Fumera, F. Roli, Multiple classifier systems for robust classifier design in adversarial environments, Int'l JMLC 1 (1) (2010) 27–41.
[77] N. Šrndić, P. Laskov, Detection of malicious pdf files based on hierarchical document structure, in: 20th NDSS, The Internet Society, 2013.
[78] H. Xu, C. Caramanis, S. Mannor, Robustness and regularization of support vector machines, JMLR 10 (2009) 1485–1510.
[79] N. Carlini, D. A. Wagner, Adversarial examples are not easily detected: Bypassing ten detection methods, in: AISec '17, ACM, 2017, pp. 3–14.
[80] X. Huang, M. Kwiatkowska, S. Wang, M. Wu, Safety verification of deep neural networks, in: 29th Int'l Conf. Computer Aided Verification, Part I, Vol. 10426 of LNCS, Springer, 2017, pp. 3–29.
[81] K. Pei, Y. Cao, J. Yang, S. Jana, Deepxplore: Automated whitebox testing of deep learning systems, in: 26th SOSP, ACM, 2017, pp. 1–18.
[82] J. Newsome, B. Karp, D. Song, Paragraph: Thwarting signature learning by training maliciously, in: RAID, LNCS, Springer, 2006, pp. 81–105.
[83] A. Barth, B. I. Rubinstein, M. Sundararajan, J. C. Mitchell, D. Song,

P. L. Bartlett, A learning-based approach to reactive security, IEEE Trans. Dependable and Sec. Comp. 9 (4) (2012) 482–493.
[84] L. I. Kuncheva, Classifier ensembles for detecting concept change in streaming data: Overview and perspectives, in: SUEMA, 2008, pp. 5–10.
[85] W. Liu, S. Chawla, Mining adversarial patterns via regularized loss minimization., Machine Learning 81 (1) (2010) 69–83.
[86] M. Großhans, C. Sawade, M. Brückner, T. Scheffer, Bayesian games for adversarial regression problems, in: 30th ICML, JMLR W&CP, Vol. 28, 2013, pp. 55–63.
[87] M. Wooldridge, Does game theory work?, IEEE IS 27 (6) (2012) 76–80.
[88] G. Cybenko, C. E. Landwehr, Security analytics and measurements, IEEE Security & Privacy 10 (3) (2012) 5–8.
[89] M. A. Torkamani, D. Lowd, On robustness and regularization of structural support vector machines, in: ICML, Vol. 32 of PMLR, 2014, pp. 577–585.
[90] C. Lyu, K. Huang, H.-N. Liang, A unified gradient regularization family for adversarial examples, in: ICDM, Vol. 00, IEEE CS, 2015, pp. 301–309.
[91] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, ArXiv e-prints.
[92] B. Biggio, I. Corona, Z.-M. He, P. Chan, G. Giacinto, D. Yeung, F. Roli, One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time, in: MCS, Vol. 9132 of LNCS, Springer, 2015, pp. 168–180.
[93] A. Bendale, T. E. Boult, Towards open set deep networks, in: IEEE CVPR, 2016, pp. 1563–1572.
[94] B. Biggio, I. Corona, G. Fumera, G. Giacinto, F. Roli, Bagging classifiers for fighting poisoning attacks in adversarial classification tasks, in: MCS, Vol. 6713 of LNCS, Springer-Verlag, 2011, pp. 350–359.
[95] D. Maiorca, B. Biggio, M. E. Chiappe, G. Giacinto, Adversarial detection of flash malware: Limitations and open issues, CoRR abs/1710.10225.
[96] B. Nelson, B. Biggio, P. Laskov, Understanding the risk factors of learning in adversarial environments, in: AISec '11, 2011, pp. 87–92.
[97] G. F. Cretu, A. Stavrou, M. E. Locasto, S. J. Stolfo, A. D. Keromytis, Casting out demons: Sanitizing training data for anomaly sensors, in: IEEE Symp. Security and Privacy, IEEE CS, 2008, pp. 81–95.
[98] C. Liu, B. Li, Y. Vorobeychik, A. Oprea, Robust linear regression against training data poisoning, in: AISec '17, ACM, 2017, pp. 91–102.
[99] J. Steinhardt, P. W. Koh, P. Liang, Certified defenses for data poisoning attacks, in: NIPS, 2017.
[100] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, N. Taft, Learning in a large function space: Privacy-preserving mechanisms for SVM learning, J. Privacy and Conf. 4 (1) (2012) 65–100.
[101] T. Dietterich, Steps Toward Robust Artificial Intelligence, AI Magazine 38 (3), 2017.
[102] Z. Lipton, The mythos of model interpretability, ICML Workshop on Human Interpretability of Machine Learning, 2016.