Position: Certified Robustness Does Not (Yet) Imply Model Security

Andrew C. Cullen [1] **Paul Montague** [2] **Sarah Erfani** [1] **Benjamin I. P. Rubinstein** [1]

Abstract

While certified robustness is widely promoted as a solution to adversarial examples in Artificial Intelligence systems, significant challenges remain before these techniques can be meaningfully deployed in real-world applications. We identify critical gaps in current research, including the paradox of detection without distinction, the lack of clear criteria for practitioners to evaluate certification schemes, and the potential security risks arising from users' expectations surrounding "guaranteed" robustness claims. These create an alignment issue between how certifications are presented and perceived, relative to their actual capabilities. This position paper is a call to arms for the certification research community, proposing concrete steps to address these fundamental challenges and advance the field toward practical applicability.

1. Introduction

A known property of learned models like neural networks is that they can have their outputs changed through semantically indistinguishable changes to their inputs (Biggio et al., 2013; Szegedy et al., 2014). The risk associated with these manipulated samples—known as *adversarial examples*—is heightened by both the confidence ascribed by models to these samples (Kumar et al., 2020), and how simple they are to construct, with mechanisms typically relying upon the same gradient descent style processes that are used in model training (Papernot et al., 2017; Carlini & Wagner, 2017b). As Artificial Intelligence (AI) increasingly permeates both interpersonal and business interactions, these adversarial examples have the potential to impact the security of real world systems (Ibitoye et al., 2019; Finlayson et al., 2019; Albert et al., 2020; Liu et al., 2025a).

[1]School of Computing and Information Systems, University of Melbourne, Australia [2]DST Group, Adelaide, Australia. Correspondence to: Andrew C. Cullen <andrew.cullen@unimelb.edu.au>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

In response to these security concerns, significant research effort has been devoted to what are known as *adversarial defenses*, which are designed to minimize or mitigate specific attacks (Chakraborty et al., 2018). However, it is crucial to note that these defenses are rarely more than *best response* strategies to particular attack vectors, as deployed mitigations can often be defeated by identifying a single, undefended vector (Perolat et al., 2018). This makes AI security into an expensive, reactive process that requires constant vigilance by model deployers.

Certified defenses eschew this best response paradigm by guaranteeing the absence of a family of potential attacks, rather than any one attack (Lecuyer et al., 2019; Li et al., 2019). These families are typically parameterized by ℓ_p inputs, with the certification guaranteeing the absence of adversarial examples within a calculated region. While the literature has primarily focused upon classifiers, recent works have begun to explore extensions of certified defenses to other problem spaces, including reinforcement learning and regression (Han et al., 2018; Lütjens et al., 2020; Hammoudeh & Lowd, 2022; Liu et al., 2023; Rekavandi et al., 2024; Liu et al., 2025b).

For all the promise of such systems, when it comes to the practical implications of these approaches, the devil is very much in the details. These techniques are presented as producing the distance to the *nearest adversarial example*, which implies that the sample being certified is itself not already attacked. However, as we observed by Cullen et al. (2024b), a more precise framing is that: a certification bounds the distance to the *nearest class changing example*. This distinction may appear minor, however, it is crucial for understanding the limitations of these techniques, as there is no guarantee that the class prediction is accurate. As such the certification could be the distance from either a clean sample to an adversarial instance, an adversarial instance to a different adversarial class, or an adversarial sample to a clean sample. Thus current certifications provide no information for distinguishing between clean and attacked samples, with certifications existing for both, and thus, there is no inherent security inferred by the certificate. Even the idea that a certification can be considered a measure for how much effort would be required to attack a particular sample does not hold, given the observation that certifications themselves can be exploited to guide adversarial

attackers (Cullen et al., 2024b).

This then leads to a problematic alignment issue, in which there is a contradiction between how these systems are presented—as reliable, *guarantees* against adversarial manipulation—and the practical reality of how they perform. And this difference is crucial for practical security, as it produces an attack surface between how the security of these systems may be perceived, relative to what it can deliver. It is for these reasons that we take the position that today's certified guarantees may provide more security theater than actual security, especially if they are providing a false-sense of security to users who are not fully across these technical nuances.

1.1. Why This Position Paper?

Adversarial examples present a clear and present danger to AI models, and the risks associated with their existence will only grow as models are more frequently integrated into systems where incentives exist for adversarial manipulation. While certified defenses have been presented as an incorruptible solution to adversarial examples, the current literature does not support their practical application. **This paper argues that current research into certified robustness is not aligned with the provision of model security, and may, in fact, harm security**. Our interrogation of this point is supported by:

1. Examining the gap between the *ideal* of certifications and their practical implications.
2. Presenting best practice for aligning research, development, and deployment to tighten model security.
3. Arguing that an application-driven approach is crucial for enhancing the impact of certifications.

2. Securing Against Adversarial Manipulation

The performance dividends made possible by deploying AI has lead to its inclusion in a broad swathe of real world systems. However, these systems introduce new frontiers of risk, as these models are incredibly sensitive to being exploited by a motivated attacker.

Within the AI security community, significant research interest has been placed upon norm-minimising ℓ_p evasion attacks (Papernot et al., 2016a), which attempt to induce a class change at test (or inference) time to minimize the ℓ_p distance between the original sample and its corresponding adversarial example. The appeal of this threat model is multifaceted: it affects a broad spectrum of systems (including classifiers and reinforcement learning); leads to easy-to-construct attacks; and correlates with our conceptual understanding of human- and machine-perceptibility (Gilmer et al., 2018).

Attacks in the form of adversarial examples can be considered as variants of gradient descent, involving finding a class flipping example $\mathbf{x}'$ that approximates the minima of

$$\begin{aligned} &\arg\min_{\mathbf{x}' \in \mathcal{S}} \|\mathbf{x} - \mathbf{x}'\|_p \\ &\text{s.t. } \arg\max_{i \in \mathcal{K}} f_i(\mathbf{x}) \neq \arg\max_{i \in \mathcal{K}} f_i(\mathbf{x}') \ , \end{aligned} \tag{1}$$

across some permissible space $\mathcal{S}$, which is typically the d-dimensional space $[0, 1]^d$ for computer vision. This framing has led to a number of distinct evasion attacks, including PGD (Madry et al., 2018), Carlini-Wagner (Carlini & Wagner, 2017b), DeepFool (Moosavi-Dezfooli et al., 2016), and AutoAttack (Croce & Hein, 2020), and has been shown to have the potential to compromise real world systems (Wu et al., 2020; Cullen et al., 2023). Similar mechanisms also can be deployed to attack models at training time, to either corrupt learning performance or embed deleterious behaviors into the model's outputs.

2.1. Adversarial Defenses

While early works suggested that techniques like model regularisation and weight decay may minimize the success rates of these attacks (Kukačka et al., 2017), these mitigations have been broadly shown to be ineffective (Kurakin et al., 2020; Athalye et al., 2018). Consequently, research has shifted to developing countermeasures, known as adversarial defenses. While these approaches have demonstrated more success, they also share a common weakness, in that they serve as responses to specific attacks, and do not typically provide resistance against alternative attack frameworks. This has led to a cyclical development process, where defenses are attacked, and new defenses are subsequently proposed to counter those attacks. An example of this is single step-attacks (Goodfellow et al., 2015) being mitigated by adversarial training, which led to the development of multi-stage attacks (Kurakin et al., 2020). These were in turn countered by defensive distillation (Papernot et al., 2016b), which has subsequently been attacked. This game of cat-and-mouse demonstrates that an attacker only needs to find an undefended vector to carry out their attack. Therefore, the adversarial resistance offered by a defense is, at best, limited when faced with a motivated attacker who can evade or exploit the deployed system (Meng & Chen, 2017; Carlini & Wagner, 2017a).

2.2. Certified Defenses

In response to the inadequacy of adversarial defenses, the AI and security communities have developed certified defenses, which construct regions around samples in which it can be guaranteed that no adversarial example exists. Crucially, these guarantees are independent of the particular attack framework, and only make basic assumptions regarding the

threat model associated with the attacker.

Certification mechanisms eschew the reactive view of adversarial defenses in favor of proactively bounding the space within which adversarial examples can exist. In some mechanisms, this might be a $\mathbf{x}$-centered p-norm ball of radius r defined as $B_p(\mathbf{x}, r)$, where r is strictly less than

$$r^\star = \inf \{\|\mathbf{x} - \mathbf{x}'\|_p \colon \mathbf{x}' \in \mathcal{S}, F(\mathbf{x}) \neq F(\mathbf{x}'), \} \quad (2)$$
$$\text{where } F(\cdot) = \mathbb{1}\left(\arg\max_{i \in \mathcal{K}} f_i(\cdot)\right) \ .$$

Here $\mathbb{1}$ is a one-hot encoding of the predicted class in $\mathcal{K} = \{1, \ldots, K\}$. The size of $B_p(\mathbf{x}, r)$ can be considered a reliable proxy for both the *detectability* of adversarial examples (Gilmer et al., 2018) and the *cost* to the attacker (Huang et al., 2011).

The construction of such bounds are typically approached through either exact or high-probability methods, with interval bound propagation (IBP) and convex relaxation (Mirman et al., 2018; Weng et al., 2018) being examples of the former, and randomized smoothing (Lecuyer et al., 2019) thelatter. While high probability methods construct high-probability bounds on the existence of adversarial examples, exact methods construct bounds by propagating intervals through the model and tracing potential class changes.

Exact approaches require significant changes to training processes and place limits on available model architectures. Moreover, these techniques impose a significant computational cost (in terms of time and GPU memory) that scales with model size, due to the need to propagate bounds through each layer (Zhang et al., 2018b; Xu et al., 2020), which typically requires the introduction of approximations to scale to model sizes of academic/industrial interest (Gowal et al., 2018; Singh et al., 2018). In contrast, randomized smoothing can be applied to almost any model architecture or training routine, with the only required changes occurring before the input layer and after the output layer. While randomized smoothing does require significant numbers of model samples to be evaluated, the computational time implications of this can be partially ameliorated, as the sampling process is embarrassingly parallel, making it a powerful alternative to bound-propagation style approaches (Cohen et al., 2019).

We must emphasize that the following definitions, and their associated techniques, are heavily aligned with evasion attacks in the interests of clarity. While some recent works have begun to consider constructing certifications against other attack frameworks, there still exist a broad range of attacks outside the aegis of evasion attacks, including backdoor attacks that embed deleterious behaviors that can be manipulated; model stealing attacks, where proprietary information is extracted from the model; check fraud, which forces the model to read a larger amount of data than what is written (Papernot et al., 2016b) and more.

2.2.1. Randomized Smoothing

The certifications constructed by randomized smoothing (Lecuyer et al., 2019) are built around a Monte Carlo estimator of the expectation of a class prediction, where

$$\frac{1}{N}\sum_{j=1}^{N} F(\mathbf{X}_j) \quad \approx \quad \mathbb{E}_{\mathbf{X}}[F(\mathbf{X})] \qquad \forall i \in \mathcal{K} \quad (3)$$
$$\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{X} \quad \overset{i.i.d.}{\sim} \quad \mathbf{x} + \mathcal{N}(0, \sigma^2) \ .$$

These expectations can be employed to provide guarantees of invariance under *additive* perturbations. In forming this aggregated classification, the model is re-construed as a *smoothed classifier*, which in turn is certified. Mechanisms for constructing such certifications include differential privacy (Lecuyer et al., 2019; Dwork et al., 2006), Rényi divergence (Li et al., 2019), and parameterising worst-case behaviors (Cohen et al., 2019; Salman et al., 2019a; Cullen et al., 2022). The latter of these approaches has proved the most performant, and yields certifications of the form

$$r = \frac{\sigma}{2}\left(\Phi^{-1}\left(\check{E}_0[\mathbf{x}]\right) - \Phi^{-1}\left(\hat{E}_1[\mathbf{x}]\right)\right) \ , \quad (4)$$

where Φ^{-1} is the inverse normal CDF, $(E_0, E_1) = \text{topk}\left(\{\mathbb{E}_{\mathbf{X}}[F(\mathbf{X})]\}, 2\right)$, and $(\check{E}_0, \hat{E}_1)$ are the lower and upper confidence bounds of these quantities to some confidence level α (Goodman, 1965).

2.2.2. Interval Bound Propagation

Conservative certificates upon the impact of norm-bounded perturbations can be constructed by way of either interval bound propagation (IBP) which propagates interval bounds through the model; or convex relaxation, which utilizes linear relaxation to construct bounding output polytopes over input bounded perturbations. In contrast to randomized smoothing, which constructs isotropic measures of ℓ_p-robustness, interval bound propagation and its associated techniques attempt to propagate the potential influence of all possible perturbations through the model, producing an anisotropic measure of the potential response of a model to any potential perturbation (Salman et al., 2019b; Mirman et al., 2018; Weng et al., 2018; Zhang et al., 2018a;b; Singh et al., 2019; Mohapatra et al., 2020). Of these, IBP is more general, while convex relaxation typically provides tighter bounds (Lyu et al., 2021).

Utilizing these techniques requires introducing an augmented loss function during training to promote tight output bounds (Xu et al., 2020). These schemes have also, until very recently, been heavily limited in the types of network architectures that they can successfully construct bounds

through, with only recent works demonstrating an applicability to a nonlinear activation functions beyond ReLU (Shi et al., 2023). Moreover they both exhibit a time and memory complexity that makes them infeasible for complex model architectures or high-dimensional data (Wang et al., 2021; Chiang et al., 2020; Levine & Feizi, 2022).

2.2.3. Global Lipschitz

Global Lipschitz takes an alternative approach to constructing certifications, a point that they distinguish through the framing of local and global robustness. The guarantees provided by prior works, which can take the form

$$\|\mathbf{x} - \mathbf{x}'\|_p \leq \epsilon \implies F(\mathbf{x}) = F(\mathbf{x}') \tag{5}$$

are considered to be local properties, that relate $\mathbf{x}$ and ϵ. Lipschitz based techniques instead attempt to construct their certifications in terms of *global* robustness, where

$$\forall \mathbf{x}_1, \mathbf{x}_2 : \|\mathbf{x}_1 - \mathbf{x}_2\|_p \leq \epsilon \implies F(\mathbf{x}_1) \stackrel{\perp}{=} F(\mathbf{x}_2) \ . \tag{6}$$

Here $\perp$ is the marker for an *abstained* class prediction, and $c_1 \stackrel{\perp}{=} c_2$ denotes that either $c_1 = \perp$, $c_2 = \perp$, or $c_1 = c_2$. In essence such a form of certification involves constructing a model that has not only an infinitesimally thin decision boundary, but a margin between the regions associated with each class, where ϵ then becomes the shortest ℓ_p distance to span the boundary. Several attempts have been made to use Lipschitz bounds during training to promote robustness. These include constructing provable lower bounds on the norm of the input manipulation required to change classifier decisions based upon the network architecture (Hein & Andriushchenko, 2017); modifying the loss associated with logits different than the ground-truth class (Tsuzuku et al., 2018); and GloRoNets, which add an additional logit corresponding to the predicted class at a point (Leino et al., 2021). While these techniques can be an order of magnitude faster than randomized smoothing, they are both less flexible—in terms of the architectures they support—and often produce smaller certifications than randomized smoothing. (Leino et al., 2021).

3. Certifications for Model Security

Whether presented as a 'certified guarantee' or a 'certified defense', that it is a 'certification' heavily implies an absolute improvement to model security. This impression is driven not just by the naming of these techniques, but also how they are described. After all, these are techniques that are guaranteed to apply in all circumstances, irrespective of the attacker's behavior. To both academic and non-academic readers who are even passingly familiar with the security risks associated with adversarial examples, such properties are incredibly appealing.

However, it must be emphasized that certified defenses do not operate in the same manner as a traditional defense. While a traditional defense ideally increases the difficulty of performing an attack, a certification only measures the distance to the nearest class-flipping example. In the literature this is typically framed as the distance to the nearest possible adversarial example, however this is not strictly true for deployed models, as *adversarial examples can also be certified* (Cullen et al., 2024b). That a certification is the distance to the nearest possible adversarial example is only true under the settings of many academic papers, in which oracle level knowledge of the true class is presumed.

This clear disparity between how certifications may be perceived, and what they actually produce presents a security risk that can potentially be exploited by motivated attackers. After all, if a model deployer is confident that their model is certifiably robust against adversarial examples, there is potentially no need to implement any other security measures. This is especially worrisome when certification mechanisms are inherently limited to specific types of threat models—for example, geometric attacks (rotational or translational) are unlikely to be covered by traditional ℓ_p based threat models (Xiao et al., 2018; Dumont et al., 2018).

In practice, as the following theorem argues, the only information provided by the certificate is the distance to the nearest potential class-flipping example, rather than providing any information regarding if the sample has been attacked or not. If a point is correctly predicted, then this distance may be the distance to the nearest adversarial example, or to the true semantic class boundary. However, if the point is an adversarial example whose class expectation is large enough to produce a certification, then the certification is the distance to the true class.

If we know that any potential attacker is ϵ bounded within the ℓ_p norm that we have been able to certify, then the guarantee will ensure that the class prediction will remain constant for these attacks. However, this does not guarantee that the prediction is correct, nor that it has not been the subject of an attack. While it may be true that certifiable adversarial examples may produce smaller certifications, due to the inherent proximity of adversarial examples to decision boundaries, this is only a heuristic, with no theoretical backing. While this observation may allow certifications to be used to stack-rank risk using certifications in a comparative fashion, we would argue that the only reliable , actionable information that a certification technique may currently provide is the absence of a certification.

Theorem 1. *A certification of size ϵ associated with the input $\mathbf{x}$ to a model f could correspond to either a certification of the correct class, that is representative of the semantic space that the sample exists within; or a certification of an incorrect class, one which is not representative of the*

semantic space a sample exists within. Thus the existence of a certification does not intrinsically provide any information regarding if the sample $\mathbf{x}$ *has been attacked or not.*

As an example of this, consider a stochastic, location invariant classifier, that produces a fixed expectation of 0.75 and a constant class prediction across all $\mathbf{x} \in \mathcal{S}$. While this classifier will certify all points, the classifier will have low accuracy, and the certified will not provide any actionable information. While this point may appear obvious, it underpins the inherent contradiction between how certifications are presented—as a security guarantee—and how they operate in actuality.

An additional consequence of the above theorem is that having access to a certification provides an attacker additional information regarding where adversarial examples may or may not be (Cullen et al., 2024b). This allows an attack to be guided not just by gradients, but by knowledge of where adversarial examples can and cannot exist. Thus, if the attacker has access to the certifications, then they have an information advantage relative to an uncertified model. Thus it can be argued that *employing certification mechanisms may compromise AI security* (Cullen et al., 2024b).

3.1. Employing Certifications

Given these observations, there is a clear need for the certification research community to acknowledge the limitations inherent to certifications, and to reflect on how the framing of these techniques may drive misapprehensions about the levels of security provided. At the most basic level, this should include emphasizing that certifications should only be accessible to those who have trusted access to the model. Throughout this paper, we will explore potential research directions relating to both this and other issues through a series of open questions.

Open Question 1. *How best should certifications be employed to enhance model security?*

Based upon our discussions to this point, we can treat a certification as a heuristic measure of how likely it is that a sample may have been manipulated. However, thinking about a certification in isolation also potentially minimizes how information security is practiced.

To take a more systematic perspective, consider organizational security as a composition of operations, that could include rules, algorithmic screening, human operators, and more. If a model produces a radius, how would that information be best served by other components of that process flow? Should the information of the certification be propagated through to subsequent tasks (or even back to earlier operations)? How could a certification be incorporated into a multifaceted assessment of risk, for both individual samples and for collective sets? Can certifications be informed by measures of risk at preceding steps of the model pipeline?

These questions may seem vague, but it is crucial to think about how techniques designed for mitigating risk—as certifications are designed to do—may exist in the context of real-world risk management frameworks. Both ISO/IEC 27001:2022 and the NIST AI Risk Management Framework (Int, 2022; National Institute of Standards and Technology, 2022) treat AI systems governance as something that requires continuous, active, multifaceted risk monitoring and assessment. For organisations, conforming to such information security controls is crucial not just for managing their own risk, but for aligning with legal and auditing expectations. In the case of smaller organisations, simply recording certifications may be enough to satisfy auditing requirements, but more complicated security apparatus will require a more nuanced perspective to be taken.

The challenge with attempting to answer questions like these in an academic context is that they do not align well with the tools that we have at our disposal. We do not have easy access to real-world information security risk frameworks. And even if we did, any testing we performed would likely produce results that were specific to particular organizations. This is not to say that these problems are not able to be studied within an academic context. In fact, facets of this problem space can be seen in the fields of mathematical risk management, human-in-the-loop computing, human computer interaction, game theory (Zhou et al., 2019; Sun et al., 2023; Cullen et al., 2024a; Adams et al., 2025), and psychology. This suggests that working towards a more holistic view of certified robustness will require multidisciplinary research expertise.

Taking such a perspective is critical to avoid certifications becoming more security-theater than actual security. As has been noted in the differential privacy community, the inherent trade off between user privacy and utility in differentially private systems creates a tension that has the potential to lead system creators to minimize transparency. Doing this has the potential to convert privacy guarantees into advertizing material and window dressing, that provides only the appearance of positive user benefit (Khare, 2009). In response to this, recent observational studies have begun to consider both how expectations of privacy shape user habits, and how clarifying private mechanisms can induce confidence in system privacy (Xiong et al., 2020; Smart et al., 2022). If certification schemes are to be considered as similarly important for demonstrating model security, then it is important to both consider and study how the framing of these mechanisms affects user expectations.

Open Question 2. *What is required for certifications to be practically deployed for end users?*

While works examining randomized smoothing, IBP, and global Lipschitz-style certifications often highlight their

relative benefits, the level of detail provided is typically insufficient for end-users to assess whether an approach suits their needs. This is particularly true when user requirements span factors such as resource demands, ease of deployment, and certification performance on datasets relevant to their use cases. We believe it is crucial for researchers to develop a shared framework for analyzing certification schemes, offering more contextual information about their performance.

Typically, certification works allude to their employed computational resources, which is sometimes supplemented with a discussion of the total computational time required. However, in practice comparisons between the resource demands imposed by different techniques are rare, and yet these very comparisons are crucial for determining the suitability of these schemes for end users. This is especially so when production environments may not share the same bottlenecks as research systems, which may lead to differing perspectives on how computational costs would be perceived.

In practice, we contend that practitioners should better structure their comparisons in terms of the resource requirements, the computational time required, and the level of achievable parallelism. While it is simple to state this as a necessity, in practice these comparisons are complicated by the stark methodological differences between the core techniques. To take randomized smoothing as an example—the large number of draws required to construct the expectations may appear to be numerically expensive. However, in practice this task of repeated model draws is embarassingly parallel, can be split over arbitrarily many GPUs, and only requires as much memory as is required to hold the model. In contrast interval bound propagation typically only requires a single pass to establish a certification. However implementing this requires both significant amounts of computational time and GPU memory to construct the certification, which intrinsically limits the size of models that can be certified.

While there is a paucity of comparisons between the different certification frameworks, the International Verification of Neural Networks Competition's (VNN-COMP) comparisons between Interval Bound style certification mechanisms serves as a proof of concept for how these comparisons could be performed (Müller et al., 2022; Brix et al., 2023). VNN-COMP builds comparisons between the success rates and running times, while controlling for run-time related issues by providing a shared codebase and prescribed computational environments, demonstrating that it is possible to begin to construct broader comparisons. However it is crucial to emphasize that the VNN-COMP comparisons only exist for IBP style certifications, and do not consider randomized smoothing or Lipschitz approaches.

The net result of constructing more rigorous estimates on computational cost will likely require a broader set of experiments than those typically performed in certification papers—especially with regard to the impact of different model sizes. However, it is important to stress that such an analysis should not be strictly rooted in trying to demonstrate the superiority of a technique, but it should rather be focused upon delving into the properties of the technique.

In the absence of clear practices for deploying certification schemes, research on computational cost should aim not to prove the superiority of any technique, but to provide knowledge that helps practitioners decide whether to use a certification framework. While this may be challenging given publishing conventions focused on state-of-the-art improvements, it could open new opportunities for comparing different certification schemas.

Open Question 3. *How do we test certification schemes in a manner that reflects real world use cases?*

Establishing certification performance on key reference datasets like MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky et al., 2009), and the Large Scale Visual Recognition Challenge variant of ImageNet (Deng et al., 2009; Russakovsky et al., 2015) are important tools for validating research works. However, the semantic properties of these datasets, and their diversity—or, more precisely, their lack thereof—limits the ability to transfer these results to other datasets of interest. This has even been shown to extend to datasets in different contexts to those in which the datasets were originally sampled, due to geographic and cultural biases that are driven by the very mechanisms through which these datasets were originally constructed (Buolamwini & Gebru, 2018; Celis & Keswani, 2020; Karkkainen & Joo, 2021; Mandal et al., 2021). While some task-specific works have begun to consider broader views on certification datasets (Dvijotham et al., 2020; Korzh et al., 2024), there clearly exists significant space for broadening the scope of how these systems are evaluated, to better demonstrate and understand utility.

As noted by Cullen et al. (2024c), the performance properties of different certification techniques can vary based upon the distribution of points within what they describe as the simplex of potential output spaces. As it is likely that datasets of interest may not share the same properties as those employed within academic research, it is important that we broaden our appreciation of what exactly state of the art is, and how techniques can be selected to maximize utility for specific tasks.

Beyond this, while improving the size of certified guarantees will always be important, it is also crucial that users are supported with the information to contextualize said guarantees. After all, a certification with an ℓ_p size of 2 (for some p) likely does not intrinsically convey enough knowledge to understand the risk associated with a sample from an arbitrary dataset being attacked—for it may be that all samples are clustered within a distance of 2 of the sample

point, or there may not be a single other clean sample within this radius. Thus, for these systems to have real world applicability and interpretability, techniques to contextualize certification sizes are crucial.

A source of inspiration for improving the quality of testing within the certification literature is Instance Space Analysis (Smith-Miles et al., 2010; Muñoz et al., 2018), which can be used to create a representative footprint of where samples may exist. This data-driven approach allows practitioners to both quantify how much coverage a dataset provides over potential input space, but also can guide the generation of new datasets. Drawing from such approaches may be useful to better understand the factors that drive certification performance in datasets that do not resemble the community's typical reference datasets.

4. Coverage

To extend upon the preceding discussions of improving certifications to enhance the concept of model security, we now turn to more practical considerations.

Open Question 4. *How do we improve the quality of guarantees provided by certifications?*

Intuitively, the size associated with a certification is directly correlated with its applicability, with larger regions of coverage providing more general guarantees, and more security. While true, it is also important to note that current certifications are often small enough to not obviate the existence of imperceptibly small adversarial perturbations. Thus increasing the size of certifications will inherently decrease the risk of attack (Gilmer et al., 2018).

This perspective on certification size being a direct measure of risk is challenged by geometric perturbations, where the ℓ_p distance may not reflect the level of difficulty in either constructing or detecting a manipulation. Moreover, any ℓ_p certification can be negated if the attacker can operate in some space $\ell_q : p \neq q$. While there is overlap between the regions of coverage provided by differing ℓ_p spaces, the potential for shifting the attack norm may introduce opportunities for the attacker to exploit.

To understand the implications of this, it is important to remember that for an attacker to be successful, they only need to find a single working adversarial example. By contrast, a defender ideally must prevent all adversarial examples from being passed through the model. In a certification context, consider a scheme that produces an ℓ_p-norm ball of size r_p. As indicated by Theorem 2, if $q < p$ no ℓ_q-norm adversarial examples exist with size $r_q < r_p$. However, if $q > p$, *smaller* ℓ_q *norm adversarial examples may exist*! Thus mismatches between attacker and certification norms can potentially induce an unfounded sense of security. This is especially so when it is possible that the certification norm potentially does not align with the capacity for these adversarial examples to be detected.

Theorem 2. *Consider a ℓ_p-certification out to a distance of r_p. Potential adversarial attacks for an attacker operating with an ℓ_q-norm attack exist for $r_q > \min\left(d^{1/q-1/p}, 1\right) r_p$.*

Proof. When $q < p$, the two regions of certification intersect at $r_q = r_p$, and thus there exists some points in the region $r_q > r_p$ that admit potential adversarial examples.

When $q > p$, the regions of certification intersect at $\frac{r_q}{d^{1/q}} = \frac{r_p}{d^{1/p}}$, for a d-dimensional space. Thus there exists points at distances $r_q > d^{1/q-1/p} r_p$ which are not covered by the ℓ_p region of certification. □

While Theorem 2 does present a mechanism for translation from ℓ_p certifications to ℓ_q threat models, ideally we should be considering how to optimize certifications for the threat model of interest, as can be seen in recent works that have begun to generalize certifications away from ℓ_2 threat models (Yang et al., 2020; Huang et al., 2023). We should also explore moving past individual threat models to instead consider maximizing certification coverage. However, this will require us to fundamentally change how we assess certification performance. While prior works have demonstrated that it is possible to ensemble certifications (Cullen et al., 2024c), their approach was still rooted in an ℓ_2 space. Ultimately maximizing coverage may require new certification mechanisms in other ℓ_p spaces and even non-ℓ_p threat models such as edit distance for sequence classifiers (Huang et al., 2023). It may also require balancing the costs of performing multiple certifications, and the added utility provided by such a layering. This leads to an additional question, regarding how new certifications can be constructed.

Open Question 5. *How can we build certification mechanisms that can be generalized to a broader set of model types?*

To this point, while we have attempted to be general in our consideration of certifications, there has been an inherent bias towards the robustness of classifiers, and classifier-like systems. This bias reflects that of the overall certification space, which is heavily weighted towards works considering classifiers under ℓ_2-norm bounded (or ℓ_p) threat models. While ensuring classifier performance is important, there is no guarantee that the AI systems that we will most heavily rely upon in the future will be similar, nor that the risks of adversarial manipulation will be limited to such classifiers (Mangal et al., 2023). While recent works have begun to consider how certified robustness can be generalized to frameworks like reinforcement learning (Lütjens et al., 2020; Kumar et al., 2021; Mu et al., 2023; Wu et al., 2022), there

still remains significant potential for expanding the scope of problems considered through certifications.

5. Secure Development

Finally, it is critical to consider how certification techniques can be developed into secure models and certification code implementations. After all, the security guarantees provided by certifications will be for naught if they cannot be incorporated into deployed code.

Open Question 6. *How do we incentivize the development of more secure code?*

To date, research projects on certification have remained at low levels of technology readiness. While this is understandable given the level of maturity of the field, if there is to be adoption of these systems, and impact outside of research, it is clear that significant care must be taken to implement secure code. This goal would be supported by systems that are open and transparent, to both enable audits (Ding et al., 2017) and enhance user confidence in system performance. Extensible implementations would also facilitate adaptation to a range of applications.

A source of inspiration is OpenDP (OpenDP, 2020; Gaboardi et al., 2020) which has generalized custom, task-specific tools to a domain-agnostic framework for enhancing the privacy guarantees associated with data access principles. OpenDP's development process is an exemplar of how security and accessibility can be balanced through an open, application driven development process, that is built upon a rigorous code- and proof-checking process. In doing so, OpenDP has managed to rapidly gain significant buy-in from both researchers and developers (Lokna et al., 2023).

It would be possible for a similar set of processes to be employed for the development of more production-ready certification code-bases. The advantages of such an approach would not just be useful for potential deployments of certification frameworks, but would also allow for different schemes to be more readily tested against each other. Such comparisons would only benefit the development of the field. Moreover, a shared research framework could also help induce a greater sense of confidence in new works, as it would be clear that they were operating upon the same framework employed by earlier techniques.

Additional lessons can also be drawn from the specific development pathways of cryptographic and differential privacy implementations, and how these pathways have lead to new research developments. Rather than directly implementing state-of-the-art research, OpenDP and crytrographic protocols have often draw inspiration from these systems while focusing upon ensuring that they work to a set of assumptions that allows for the develop of reliably robust, testable, and verifiable systems. This focus upon how systems are employed in practice, in contrast to the more standard academic assumptions that had existed in prior works then in turn lead to new areas of research interest in areas like rounding and floating point issues in Differential Privacy correctly (Mironov, 2012; Balcer & Vadhan, 2017; Jin et al., 2022), the development of side-channel (Jin et al., 2022) and floating-point attacks (Jin et al., 2024b), and explorations of the impact of privacy budgets (Jin et al., 2024a). These examples provide a clear precedent for how examining security problems with an eye to how they will be deployed within the real world can pay both practical and research dividends.

6. Alternative Views

While the preceding content argues that certified robustness schemes share significant open weaknesses, a counterfactual perspective would be that any improvements provided by a certification still enhances model security, even if these improvements do not provide complete security. While this is true, as we have argued within this paper, we believe there is a high likelihood that these schemas will result in security theatre, rather than security. The likelihood of this is heavily driven by the presentation of certifications as a guarantee of robustness. When certifications are marketed as definitive proof of a models resilience to adversarial examples, it creates the illusion of a level of protection that may not truly exist. Such a false sense of security could divert attention from more robust, ongoing security measures and research, potentially leading users to neglect further model improvements or defensive strategies. Moreover, incorporating certifications may also foster complacency in model development.

It could also be argued that these expectations are not the responsibility of the certification community, and that these systems are being developed for technical users at this stage, with future developments taking care of ease of use and broader adoption. While basic research is undeniably important, these benefits do not negate concerns for how certifications are being communicated. For technical users, it is essential to understand the limits of certified robustness guarantees and the broader implications of these systems, particularly in industries that rely on AI models. If certifications are not accompanied by clear explanations of their limitations and scope, there is a risk that non-technical stakeholders—who may not fully grasp the underlying complexities—could misinterpret or overestimate the significance of these certifications. This disconnect could lead to misuse or overreliance on certifications, which in turn may hinder the development of more comprehensive security strategies. In the long term, as these systems become more accessible and widespread, there will be an increased need for transparency and clear communication about what

certified robustness can and cannot achieve. Without this, the potential for security theater remains a significant concern.

7. Conclusions

The AI security community has consistently been producing research that moves the needle in terms of our understanding of risks facing models, and the strategies that can be employed to mitigate said risks. However, for all of the research insights gained, there is a strong case to be made that the current literature has not yet bridged the gap from practical promise to deployable applications. This is particularly true for certified robustness, given the expectations of users searching for a guaranteed solution to the risk of adversarial manipulation.

Indeed, the very framing of certified robustness as providing guarantees of adversarial resistance can create a false sense of security, and more broadly, an alignment issue between how users would likely perceive these systems and their actual performance. Given this, within this paper we argue that additional care must be taken to ensure that these certifications are presented in a manner that prevents their being misinterpreted as blanket protections, rather than constrained assurances tied to specific threat models, perturbation bounds, and data distributions. Such misunderstanding can lead to adverse security outcomes, especially in real-world deployment scenarios where defenders do not have access to oracle information, and where adversaries do not necessarily conform to the narrow theoretical bounds underpinning certification techniques.

This work is not just an excoriation of current research practices. We also argued that considering how these systems are used and perceived can inspire new, interesting research directions. This work would extend beyond the current remit of certification research—which is broadly focused on improving bulk metrics on reference data sets—into exploring the how and why of certification performance, the factors that incentivize the development of secure code, the applicability of certifications to real world threat models, and how they should be communicated to stakeholders. These rich new veins of research questions have the potential to significantly improve the safety of deployed AI systems.

Impact Statement

This work takes the position that for all its promise, certified robustness still has a long way to go before being ready for wide-spread deployement as a real-world method of securing AI. While this position may superficially appear to be pessimistic, we believe that constructive discussion about the current state of certification research can help reveal new, productive research directions. If the research community can progress solutions, we believe that certifications can provide a tangible impact on AI security, especially where AI is deployed in high-risk and high stakes contexts, for positive societal impact.

Acknowledgments

This work was supported by the Australian Defence Science and Technology (DST) Group via the Advanced Strategic Capabilities Accelerator (ASCA) program.

References

Takuma A Adams, Andrew C. Cullen, and Tansu Alpcan. Suboptimality of Constrained Action Adversarial Cyber-Physical Games. *Dynamic Games and Applications*, pp. 1–20, 2025.

Kendra Albert, Jonathon Penney, Bruce Schneier, and Ram Shankar Siva Kumar. Politics of Adversarial Machine Learning. *arXiv preprint arXiv:2002.05648*, 2020.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pp. 274–283. PMLR, 2018.

Victor Balcer and Salil Vadhan. Differential Privacy on Finite Computers. *arXiv preprint arXiv:1709.05396*, 2017.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion Attacks Against Machine Learning at Test Time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, ECMLPKDD, pp. 387–402. Springer, 2013.

Christopher Brix, Stanley Bak, Changliu Liu, and Taylor T Johnson. The Fourth International Verification of Neural Networks Competition (VNN-COMP 2023): Summary and Results. *arXiv preprint arXiv:2312.16760*, 2023.

Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91. PMLR, 2018.

Nicholas Carlini and David Wagner. MagNet and "Efficient Defenses Against Adversarial Attacks" are not Robust to Adversarial Examples. *arXiv preprint arXiv:1711.08478*, 2017a.

Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (S & P)*, pp. 39–57. IEEE, 2017b.

L Elisa Celis and Vijay Keswani. Implicit Diversity in Image Summarization. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial Attacks and Defences: A Survey. *arXiv preprint arXiv:1810.00069*, 2018.

Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified Defenses for Adversarial Patches. In *International Conference on Learning Representations*, ICLR, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning*, ICML, pp. 1310–1320. PMLR, 2019.

Francesco Croce and Matthias Hein. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. In *International Conference on Machine Learning*, ICML, pp. 2206–2216. PMLR, 2020.

Andrew C. Cullen, Paul Montague, Shijie Liu, Sarah Monazam Erfani, and Benjamin I.P. Rubinstein. Double Bubble, Toil and Trouble: Enhancing Certified Robustness through Transitivity. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19099–19112. NeurIPS, 2022.

Andrew C. Cullen, Benjamin I.P. Rubinstein, Sithamparanathan Kandeepan, Barry Flower, and Philip HW Leong. Predicting Dynamic Spectrum Allocation: A review covering Simulation, Modelling, and Prediction. *Artificial Intelligence Review*, 56(10):10921–10959, 2023.

Andrew C. Cullen, Tansu Alpcan, and Alexander Kalloniatis. Game-Theoretic Analysis of Adversarial Decision Making in a Complex Socio-Physical System. *Dynamic Games and Applications*, pp. 1–20, 2024a.

Andrew C. Cullen, Shijie Liu, Paul Montague, Sarah M. Erfani, and Benjamin I.P. Rubinstein. Et Tu Certifications: Robustness Certificates Yield Better Adversarial Examples. In *Forty-first International Conference on Machine Learning*, 2024b.

Andrew C. Cullen, Paul Montague, Shijie Liu, Sarah Erfani, and Benjamin I. P. Rubinstein. It's Simplex! Disaggregating Measures to Improve Certified Robustness. In *2024 IEEE Symposium on Security and Privacy (SP)*, 2024c. Accepted.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 248–255. IEEE, 2009.

Zeyu Ding, Yuxin Wang, Guanhong Wang, Danfeng Zhang, and Daniel Kifer. Privacy Audits for Differential Privacy. *Proceedings on Privacy Enhancing Technologies (PoPETS)*, 2017.

Beranger Dumont, Simona Maggio, and Pablo Montalvo. Robustness of Rotation-Equivariant Networks to Adversarial Perturbations. In *ICML Workshop on "Towards learning with limited labels: Equivariance, Invariance, and Beyond"*, 2018.

Krishnamurthy Dj Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A Framework for Robustness Certification of Smoothed Classifiers using f-Divergences. In *International Conference on Learning Representations*, 2020.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography Conference*, TCC, pp. 265–284. Springer, 2006.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial Attacks on Medical Machine Learning. *Science*, 363 (6433):1287–1289, 2019.

Marco Gaboardi, Michael Hay, and Salil Vadhan. A Programming Framework for OpenDP. *Manuscript, May*, 2020.

Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the Rules of the Game for Adversarial Example Research. *arXiv preprint arXiv:1807.06732*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, ICLR, 2015.

Leo A Goodman. On Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics*, 7(2):247–254, 1965.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable Certified Robustness via Interval Bound Propagation. In *International Conference on Learning Representations*, 2018.

Zayd Hammoudeh and Daniel Lowd. Reducing Certified Regression to Certified Classification. *arXiv preprint arXiv:2208.13904*, 2022.

Yi Han, Benjamin IP Rubinstein, Tamas Abraham, Tansu Alpcan, Olivier De Vel, Sarah Erfani, David Hubczenko, Christopher Leckie, and Paul Montague. Reinforcement Learning for Autonomous Defence in Software-Defined Networking. In *Decision and Game Theory for Security: 9th International Conference, GameSec 2018, Seattle, WA, USA, October 29–31, 2018, Proceedings 9*, pp. 145–165. Springer, 2018.

Matthias Hein and Maksym Andriushchenko. Formal Guarantees on the Robustness of a Classifier Against Adversarial Manipulation. In *Advances in Neural Information Processing Systems*, volume 30 of *NeurIPS*, 2017.

Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, pp. 43–58, 2011.

Zhuoqun Huang, Neil G Marchant, Keane Lucas, Lujo Bauer, Olga Ohrimenko, and Benjamin I. P. Rubinstein. RS-Del: Edit distance robustness certificates for sequence classifiers via randomized deletion. In *Advances in Neural Information Processing Systems*, NeurIPS, pp. 18676–18711, 2023.

Olakunle Ibitoye, Rana Abou-Khamis, Ashraf Matrawy, and M Omair Shafiq. The Threat of Adversarial Attacks on Machine Learning in Network Security–A Survey. *arXiv preprint arXiv:1911.02621*, 2019.

ISO/IEC 27001:2022: Information security, cybersecurity and privacy protection Information security management systems Requirements. International Organization for Standardization and International Electrotechnical Commission, Geneva, Switzerland, 2022. Available at: https://www.iso.org/standard/82875.html.

Jiankai Jin, Eleanor McMurtry, Benjamin IP Rubinstein, and Olga Ohrimenko. Are We There Yet? Timing and Floating-Point Attacks on Differential Privacy Systems. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 473–488. IEEE, 2022.

Jiankai Jin, Chitchanok Chuengsatiansup, Toby Murray, Benjamin IP Rubinstein, Yuval Yarom, and Olga Ohrimenko. Elephants Do Not Forget: Differential Privacy with State Continuity for Privacy Budget. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1909–1923, 2024a.

Jiankai Jin, Olga Ohrimenko, and Benjamin I. P. Rubinstein. Getting a-Round Guarantees: Floating-Point Attacks on Certified Robustness. In *Proceedings of the 17th ACM Workshop on Security and Artificial Intelligence*, AISec, 2024b. accepted.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021.

Rohit Khare. Privacy Theater: Why Social Networks Only Pretend to Protect You. https://techcrunch.com/2009/12/27/privacy-theater/, 2009. TechCrunch (Online); accessed 7 January 2025.

Dmitrii Korzh, Elvir Karimov, Mikhail Pautov, Oleg Y Rogov, and Ivan Oseledets. Certification of Speaker Recognition Models to Additive Perturbations. *arXiv preprint arXiv:2404.18791*, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

Jan Kukačka, Vladimir Golkov, and Daniel Cremers. Regularization for Deep Learning: A Taxonomy. *arXiv preprint arXiv:1710.10686*, 2017.

Aounon Kumar, Alexander Levine, and Soheil Feizi. Policy Smoothing for Provably Robust Reinforcement Learning. *arXiv preprint arXiv:2106.11420*, 2021.

Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial Machine Learning-Industry Perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 69–75. IEEE, 2020.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, ICLR, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified Robustness to Adversarial Examples with Differential Privacy. In *2019 IEEE Symposium on Security and Privacy (S & P)*, pp. 656–672. IEEE, 2019.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-Robust Nneural Networks. In *International Conference on Machine Learning*, pp. 6212–6222. PMLR, 2021.

Alexander Levine and Soheil Feizi. (de)Randomized Smoothing for Certifiable Defense against Patch Attacks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6465–6475. NeurIPS, 2022.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified Adversarial Robustness with Additive Noise. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9459–9469. NeurIPS, 2019.

Shijie Liu, Andrew C. Cullen, Paul Montague, Sarah M. Erfani, and Benjamin I.P. Rubinstein. Enhancing the Antidote: Improved Pointwise Certifications against Poisoning Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8861–8869, 2023.

Shijie Liu, Andrew C. Cullen, Paul Montague, Sarah M. Erfani, and Benjamin I.P. Rubinstein. Multi-level Certified Defense against Poisoning Attacks in Offline Reinforcement Learning. In *International Conference on Learning Representations*, ICLR, 2025a.

Shijie Liu, Andrew C. Cullen, Paul Montague, Sarah Monazam Erfani, and Benjamin IP Rubinstein. Fox in the Henhouse: Supply-Chain Backdoor Attacks against Reinforcement Learning. *arXiv preprint arXiv:2505.19532*, 2025b.

Johan Lokna, Anouk Paradis, Dimitar I Dimitrov, and Martin Vechev. Group and Attack: Auditing Differential Privacy. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1905–1918, 2023.

Björn Lütjens, Michael Everett, and Jonathan P How. Certified Adversarial Robustness for Deep Reinforcement Learning. In *Conference on Robot Learning*, pp. 1328–1337. PMLR, 2020.

Zhaoyang Lyu, Minghao Guo, Tong Wu, Guodong Xu, Kehuan Zhang, and Dahua Lin. Towards Evaluating and Training Rerifiably Robust Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 4308–4317, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, ICLR, 2018.

Abhishek Mandal, Susan Leavy, and Suzanne Little. Dataset Diversity: Measuring and Mitigating Geographical Bias in Image Search and Retrieval. In *Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing*, pp. 19–25, 2021.

Ravi Mangal, Klas Leino, Zifan Wang, Kai Hu, Weicheng Yu, Corina Pasareanu, Anupam Datta, and Matt Fredrikson. Is Certifying ℓ_p Robustness Still Worthwhile? *arXiv preprint arXiv:2310.09361*, 2023.

Dongyu Meng and Hao Chen. MagNet: A Two-Pronged Defense Against Adversarial Examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 135–147, 2017.

Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable Abstract Interpretation for Provably Robust Neural Networks. In *International Conference on Machine Learning*, ICML, pp. 3578–3586. PMLR, 2018.

Ilya Mironov. On Significance of the Least Significant Bits for Differential Privacy. In *Proceedings of the 2012 ACM conference on Computer and Communications Security*, pp. 650–661, 2012.

Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Towards Verifying Robustness of Neural Networks Against A Family of Semantic Perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 244–252, 2020.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 2574–2582, 2016.

Ronghui Mu, Wenjie Ruan, Leandro Soriano Marcolino, Gaojie Jin, and Qiang Ni. Certified Policy Smoothing for Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15046–15054, 2023.

Mark Niklas Müller, Christopher Brix, Stanley Bak, Changliu Liu, and Taylor T Johnson. The Third International Verification of Neural Networks Competition (VNN-COMP 2022): Summary and Results. *arXiv preprint arXiv:2212.10376*, 2022.

Mario A Muñoz, Laura Villanova, Davaatseren Baatar, and Kate Smith-Miles. Instance Spaces for Machine Learning Classification. *Machine Learning*, 107:109–147, 2018.

National Institute of Standards and Technology. *AI Risk Management Framework (AI RMF) 1.0*. Gaithersburg, MD, USA, 2022. Available at: https://www.nist.gov/itl/ai-risk-management-framework.

OpenDP. The OpenDP White Paper. Technical report, OpenDP, 2020.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387. IEEE, 2016a.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (S & P)*, pp. 582–597. IEEE, 2016b.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks against Machine Learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519, 2017.

Julien Perolat, Mateusz Malinowski, Bilal Piot, and Olivier Pietquin. Playing the Game of Universal Adversarial Perturbations. *arXiv preprint arXiv:1809.07802*, 2018.

Aref Miri Rekavandi, Farhad Farokhi, Olga Ohrimenko, and Benjamin I. P. Rubinstein. Certified adversarial robustness via randomized α-smoothing for regression models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, NeurIPS, 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11292–11303. NeurIPS, 2019a.

Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A Convex Relaxation Barrier to Tight Robustness Verification of Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9835–9846. NeurIPS, 2019b.

Zhouxing Shi, Qirui Jin, Huan Zhang, Zico Kolter, Suman Jana, and Cho-Jui Hsieh. Formal Verification for Neural Networks with General Nonlinearities via Branch-And-Bound. In *2nd Workshop on Formal Verification of Machine Learning (WFVML 2023)*, 2023.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and Effective Robustness Certification. In *Advances in Neural Information Processing Systems*, NeurIPS, 2018.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An Abstract Domain for Certifying Neural Networks. *Proceedings of the ACM on Programming Languages*, 3(POPL):1–30, 2019.

Mary Anne Smart, Dhruv Sood, and Kristen Vaccaro. Understanding Risks of Privacy Theater with Differential Privacy. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24, 2022.

Kate Smith-Miles, Jano Van Hemert, and Xin Yu Lim. Understanding TSP Difficulty by Learning from Evolved Instances. In *Learning and Intelligent Optimization: 4th International Conference, LION 4, Venice, Italy, January 18-22, 2010. Selected Papers 4*, pp. 266–280. Springer, 2010.

Guoxin Sun, Tansu Alpcan, Seyit Camtepe, Andrew C. Cullen, and Benjamin I.P. Rubinstein. An Adversarial Strategic Game for Machine Learning as a Service using System Features. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 2508–2510, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations*, ICLR, 2014.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. NeurIPS, 2018.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-CROWN: Efficient Bound Propagation with Per-Neuron Split Constraints for Neural Network Robustness Verification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29909–29921. NeurIPS, 2021.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards Fast Computation of Certified Robustness for ReLU Networks. In *International Conference on Machine Learning*, ICML, pp. 5276–5285. PMLR, 2018.

Fan Wu, Linyi Li, Chejian Xu, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, and Bo Li. COPA: Certifying Robust Policies for Offline Reinforcement Learning against Poisoning Attacks. *arXiv preprint arXiv:2203.08398*, 2022.

Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 1–17. Springer, 2020.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially Transformed Adversarial Examples. In *International Conference on Learning Representations*, 2018.

Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. Towards Effective Differential Privacy Communication for Users Data Sharing Decision and Comprehension. In *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 392–410. IEEE, 2020.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1129–1141. NeurIPS, 2020.

Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized Smoothing of All Shapes and Sizes. In *International Conference on Machine Learning*, pp. 10693–10705. PMLR, 2020.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient Neural Network Robustness Certification with General Activation Functions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 4939–4948. Curran Associates, Inc., 2018a.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient Neural Network Robustness Certification with General Activation Functions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 4939–4948. NeurIPS, 2018b.

Yan Zhou, Murat Kantarcioglu, and Bowei Xi. A Survey of Game Theoretic Approach for Adversarial Machine Learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1259, 2019.