Doppelgängers and Adversarial Vulnerability

George Kamberov[*]
University of Alaska Anchorage
Anchorage, AK 99508
gkamberov@alaska.edu

Abstract

Machine learning (ML) classifiers can make mistakes that are perceptually and cognitively disturbing to humans. The most notorious examples of such errors are adversarial visual metamers. This paper investigates the phenomenon of adversarial Doppelgängers (AD), which encompasses adversarial visual metamers, and compares the performance and robustness of ML classifiers to human performance.

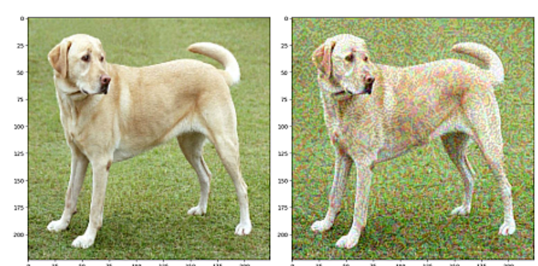
We find that ADs are inputs that are close to each other with respect to a perceptual metric defined in this paper, and show that ADs are qualitatively different from the usual adversarial examples. The vast majority of classifiers are vulnerable to ADs and robustness-accuracy trade-offs may not improve them. Some classification problems do not admit any AD-robust classifiers because the underlying classes are ambiguous. We provide criteria to determine whether a classification problem is well defined; describe the structure and attributes of AD-robust classifiers; introduce and explore the notions of conceptual entropy and regions of conceptual ambiguity for classifiers that are vulnerable to AD attacks; and discuss methods to bound the AD fooling rate of an attack. We define the notion of classifiers that exhibit hypersensitive behavior, that is, classifiers whose only mistakes are adversarial Doppelgängers. Improving the AD robustness of hypersensitive classifiers is equivalent to improving accuracy. We identify conditions guaranteeing that all classifiers with sufficiently high accuracy are hypersensitive.

1. Introduction

Perceptual metamers[1] are the most striking adversarial examples studied by the machine learning community. Two perceptual metamers are shown in Figure 1. The phenomenon of metamersim studied in the visual domain, including perceptual metamers, is a manifestation of the existence of Doppelgängers: different inputs or stimuli that are perceptually indiscriminable. The research community has engaged in active studies of adversarial vulnerability ever since the publication of [73]. Adversarial Doppelgängers, that is, adversarial examples which are Doppelgängers, are qualitatively different from the vast majority of known adversarial examples which humans readily discriminate from correctly classified input samples (Figure 2). There is no

(a) (b)

Figure 1. Most people cannot discriminate image (a) from image (b). MobileNetV2 classifies the later image as "persian" and the former picture as "taby".



(a) Labrador (b) Weimeraner

Figure 2. Applying a Fast Signed Gradient perturbation to the image (a) classified by MobileNetV2 as Labrador yields the image (b) which is classified by MobileNetV2 as Weimeraner.

*This material is based upon work supported by the National Science Foundation under Grant No. 2433241

[1]"images that are physically distinct but perceptually indistinguishable", [6]. See also "metameric images", [37]

evidence that Doppelgängers can be studied and understood completely using the ℓ_p norms or more general geodesic distances on manifolds, which have been employed to quantify sample differences and to investigate adversarial examples. Perception and context impose topology on a space of inputs, but it rarely aligns with a manifold topology. We will denote this context-relative topology by $\tau_{\eth}$. It is defined by the context-relative ability to acquire and deploy knowledge.

In this paper, we explore the context-relative perceptual topologies on a space of inputs $\mathbf{X}$ and examine the vulnerability and robustness of machine learning classifiers to adversarial Doppelgängers. We show that, while the majority of classifiers are vulnerable to adversarial Doppelgängers, safe (Doppelgänger robust) classifiers do exist if the classification problem is well defined. However, these robust classifiers may be very rare.

In Section 3, we discuss the context-relative notion of (active) indiscriminability and the topology $\tau_{\eth}$ that it induces on a space of inputs. The separability and metric properties of various motivating examples of $\tau_{\eth}$. Additionally, we show that the distances between Doppelgängers are small if measured by a perceptually-based context-relative metric. We also examine the relation between indiscernability, indsicriminability, and feature representations. The existence and structure of Doppelgänger robust classifiers is discussed in Section 4. In Section 5, we investigate the relationship between Doppelgängers and misclassified input samples, define the notion of hypersensitive behavior, and show that improved adversarial Doppelgänger robustness does not have to lead to a reduction in accuracy.

The structure of perceptually regular, i.e., AD robust classifiers are discussed in Section 6. By definition, a classifier is not regular if and only if some inputs can be attacked by adversarial Doppelgängers. Not surprisingly, it turns out that some inputs are more vulnerable than others. In Section 7, we provide measures of adversarial Doppelgängers vulnerability and upper bounds on the fooling rate of an adversarial Doppelgänger attack.

2. Related work

The pair $(\mathbf{X}, \text{indiscriminability relation})$ is a tolerance space. Tolerance spaces, rough sets, and granular computing have been discussed extensively. See [59, 62, 66, 88, 89]. The color and image metamers studied by many authors including [3, 6, 9, 18, 22, 37, 39, 41, 42, 51, 76, 86] are Doppelgängers. [2]

The research on adversarial examples to date builds on the hypothesis that the space of input samples is a metric space $(\mathbf{X}, \text{dist}_{\mathbf{X}})$. A misclassified input x^* is considered an adversarial example if it is nearby a correctly classified input sample x, i.e., $\text{dist}_{\mathbf{X}}(x, x^*)$ is small. Usually $\mathbf{X}$ is assumed to be $\mathbb{R}^n$, endowed with the ℓ_p, norm, $p = 1, 2, \ldots, \infty$ or at least locally homeomorphic to $\mathbb{R}^n$, i.e., a manifold, equipped with some geodesic distance. Somewhat non surprisingly many authors have shown that every classifier can be attacked with such adversarial examples [5, 17, 25, 52] or at least that this is true in many contexts, [44, 47, 67].

Other papers indicate that there are paths toward eliminating adversarial examples completely, i.e., it is possible to achieve provable "adversarial robustness" by fixing/retraining the classifier [1, 24, 36, 43, 72, 74]. A widely accepted tenet is that "there is a clear trade-off between accuracy and [adversarial] robustness, and a better performance in testing accuracy in general reduces [adversarial] robustness", [71]. For empirical evidence for this trade-off and some attempts to explain this phenomenon see [71, 78, 90].

3. Perceptual Topology

3.1 Indiscriminability and Topology

The ability to decide whether one stimulus/input is distinct from another is essential for adaptation, survival, and intelligent life. Intelligent agents are uniquely capable to activate knowledge to judge distinction. Williamson calls this context-relative process discrimination ([84]) and defines a context-relative symmetric and reflexive binary relation denoted by $\overset{\alpha\eth}{\approx}$ and called **indiscriminability**: [3]

Definition 1 ([84])**.** Two inputs x and y are called **indiscriminable** to a subject at a time t if and only if at time t the subject is not able to activate (acquire or employ) the relevant kind of knowledge that x and y are distinct.[4]

Indiscriminability generates a context-relative topology on the set of inputs $\mathbf{X}$.

Definition 2. The **the phenomenal neighborhood** of an input $x \in \mathbf{X}$, is the set $\mathfrak{d}(x) = \left\{ y \in \mathbf{X} : y \overset{\alpha\eth}{\approx} x \right\}$. A point $y \in \mathfrak{d}(x) \setminus \{x\}$ is called a **Doppelgänger** [of x]. The **perceptual topology** $\tau_{\eth}$ is the topology generated by the sub-basis $\mathfrak{D}_{\alpha\delta} = \{\mathfrak{d}(x)\}_{x \in \mathbf{X}}$.

Example 0: An input $x \in \mathbf{X}$ is called **optimal** if it does not have non-trivial/non-identical Doppelgängers, i.e.,

[2] Some metamers arising in other fields including biology and chemistry are not Doppelgängers, for example, segments in many earthworms are considered metamers but are visually discriminable.

[3] Context and its role in discrimination and similarity judgments have been studied extensively and by many authors including [11, 23, 30, 45, 60, 77, 79, 81].

[4] Some authors refer to indiscriminability as **active indiscriminability**, see [16]. Poincaré discusses indiscriminability in [58] but refers to it as *indiscernability*. Similarly Poston studies *indistinguishibility*, basing it on the "limit of discrimination" of the biological senses and instruments, [59].

if $\mathfrak{d}(x) = \{x\}$.[5] In particular, if $\overset{\alpha\delta}{\approx}$ is the identity relationship $=$ (i.e,. all inputs are optimal within the given context), then the perceptual topology τ_δ is discrete. However, the finiteness of human observations (they are subject to finite time and finite work constraints) and the *bounded rationality* constraints imposed by the limitations on the availability of information and computational capabilities to humans, [68], indicate that the scenario $\mathfrak{d}(x) = \{x\}$ for every $x \in \mathbf{X}$ may be highly unlikely.

Often reflexive binary relations are defined and discussed as coverings of the underlying space. Indeed, every reflexive binary relation (RBR) $\approx$ on $\mathbf{X}$ defines a covering $\{\mathfrak{g}(x) = \{y : y \approx x\}\}_{x\in\mathbf{X}}$, of $\mathbf{X}$ and vice versa one can define reflexive binary relationships through coverings of $\mathbf{X}$. See Appendix, Section A. In particular, the perceptual topology, τ_δ, is a tolerable topology, cf., Definition 10 in Appendix, Section A.

The active psychophysics research on just noticeable difference initiated by Weber and Fechner, [19, 20, 82, 83], provides one of the few classes of examples where we have empirically supported understanding of the perceptual topology.

Example 1: Let $\mathbf{X}$ be the closed bounded interval $[a, b] \subset (0, +\infty)$. Suppose that Weber's law holds and let $k > 0$ be the Weber constant. Let $w = 1 + k$, then

$$\mathfrak{d}(x) = \begin{cases} [a, xw)\,, & a \le x < aw \\ (x/w, xw)\,, & aw \le x \le b/w \\ (x/w, b]\,, & b/w < x \le b. \end{cases} \tag{1}$$

The covering $\{\mathfrak{d}(x)\}_{x\in[a,b]}$ defines a symmetric RBR but the relation is not transitive. The corresponding perceptual topology is T_0 but not T_1, and the topology is not pseudometric.

The transitivity or more more often the lack of transitivity of $\overset{\alpha\delta}{\approx}$ have been studied extensively and proven or postulated in many human experiences, [2, 6, 10, 16, 26, 27, 32, 54, 61, 84, 85].

Definition 3. We will denote by $\sim_\sigma$ the transitive closure of the indiscriminability relation $\overset{\alpha\delta}{\approx}$ on $\mathbf{X}$. It is defined explicitly as $x \sim_\sigma y$ iff there exists a finite chain of Doppelgängers $x = x_0 \overset{\alpha\delta}{\approx} x_1 \overset{\alpha\delta}{\approx} x_2 \overset{\alpha\delta}{\approx} \cdots \overset{\alpha\delta}{\approx} x_n = y$. We will call the relation $\sim_\sigma$ perceptual **metamorphy** and will refer to any two inputs $x \sim_\sigma y$ as **metamorphic**. Extending Pawlak's terminology, [53] we call the equivalence classes in $\mathbf{X}/\!\sim_\sigma$ (perceptually) **elementary sets**.

Example 2: If the indiscriminability relation is transitive, then each $\mathfrak{d}(x)$ is an elementary set. The perceptual topology may be optimal (recall Example 0) or not. In the former case it is Hausdorff and in fact $(\mathbf{X}, \tau_\delta)$ is a discrete manifold, in the later case the topology τ_δ is not T_0. See Part A.1 in the Appendix. Human visual perception provides a fundamental example where $\mathfrak{d}(x) \ne \{x\}$. If two images x and y differ only in unattended regions for example due to low saliency values (cf. [92]), then $x \overset{\alpha\delta}{\approx} y$. Visual metamers have been studied by many authors, including [3, 6, 9, 18, 22, 28, 37, 39, 41, 42, 51, 76, 86]. In these studies, the input space is assumed to be endowed with Grassmann structure (see [39]), and in particular, the indiscriminability relation is transitive.[6]

We are not aware of perceptual topologies that are metric. Still, every Doppelgänger $y \overset{\alpha\delta}{\approx} x$ of an input x is a small perturbation of x in the sense that the y is a nearest neighbor of x with respect to an appropriate metric $d_w(\cdot,\cdot)$ on $\mathbf{X}$. Indeed, let the **discrimination graph** $\Gamma(\mathbf{X}, E_{\alpha\delta})$ be the undirected simple graph, where $\mathbf{X}$ is the set of vertices and we say that there is an edge $\{x, y\} \in E_{\alpha\delta}$ between the vertices $x, y \in \mathbf{X}$ iff $x \overset{\alpha\delta}{\approx} y$.

Definition 4. We will call the discrimination graph distance $d_\infty(x, y)$ between the vertices x and y and, in particular, $d_\infty(x, y) = \infty$ iff $x \not\sim_\sigma y$ the **extended perceptual distance**. The **perceptual distance** is the metric $d_w : \mathbf{X} \times \mathbf{X} \to [0, 1]$ defined by:

$$d_w(x, y) = \frac{d_\infty(x, y)}{1 + d_\infty(x, y)}, \forall x, y \in \mathbf{X}. \tag{2}$$

The metric d_w does not generate the perceptual topology.[7] It is certainly not the usual l_p or any other manifold metric used in ML.[8]

3.2 Indiscriminabile may not be Indiscernible

Let Φ be the space of all features of the inputs/stimuli $x \in \mathbf{X}$ and let $\Phi_x \subset \Phi$ be the set of features attributed to x in a given context. Following [21], we say that x and y are **indiscernible**, in a given context, if $\Phi_x = \Phi_y$.[9] Many researchers use the terms indiscriminability and indiscernibility as synonyms. This is only accurate when $\overset{\alpha\delta}{\approx}$ is transi-

[5] Optimal objects and light sources have been described and studied in colorimetry, [42, 87].

[6] In these studies, indiscriminable inputs are referred to as "matching" or "metameric", or "alike".

[7] The open metric ball $\mathring{B}^w_{1/2}(x)$ equals $\{x\}$, for all inputs $x \in \mathbf{X}$. On the other hand, the finiteness of human observations and the hypothesis of bounded rationality suggest that biologically plausible perceptual topologies are not discrete.

[8] The existence of the extended metric was hinted at in [46]; it was discussed in a related context in [65] and rediscovered and exploited in [59]. For more discussion see Part F.1 in the Appendix.

[9] Leibniz discussed indiscernability and postulated the Principle/Law of Identity of Indiscernibles, $(\Phi_x = \Phi_y) \implies (x = y)$, in [80] and in the third and fourth papers addressed to Samuel Clarke, [8].

tive.[10] In general, the relationship between indiscriminability and indiscernibility is not well understood. However, indiscernibility implies indiscriminability if $\{\Phi_x\}_{x\in\mathbf{X}}$ is a perceptually **discriminative feature representation**, i.e.,

$$\Phi_x \bigcap \Phi_y \neq \emptyset \Longleftrightarrow x \overset{\alpha\delta}{\approx} y. \tag{3}$$

The biological plausibility of discriminative feature representations is an open question. Still, they provide insight into the structure of the perceptual topology. The attributed discriminative features[11] represent structures of Doppelgängers. Indeed, let $\{\Phi_x\}_{x\in\mathbf{X}}$ be a a feature representation and let $\mathfrak{cl}(\xi)$ be the context-dependent **semantic cluster** of inputs sharing the feature $\xi \in \Phi$. Specifically,

$$\mathfrak{cl}(\xi) = \{x \in \mathbf{X},\ \text{s.t.}, \xi \in \Phi_x\}.^{12} \tag{4}$$

In particular, if $\{\Phi_x\}_{x\in\mathbf{X}}$ is a discriminative feature representation, then every attributed discriminative feature $\xi \in \Phi_x$ is associated to, and in some way explained by, a collection of Doppelgängers since $\mathfrak{cl}(\xi) \subset \mathfrak{d}(x)$. For more detailed discussion and examples of discriminative feature representations see Part B in the Appendix.

4. Classifiers and Adversarial Doppelgängers.

A classifier R (with m labels) is called **fully populated** iff the labeling function $\text{label}_R : \mathbf{X} \to \{1, \ldots, m\}$ is surjective mapping onto the range of labels $\{1, \ldots, m\}$. For any classifier R, with m labels, we will denote by R_c the level set of the labeling function label_R for each label $c \in \{1, \ldots, m\}$.[13]

Definition 5. We say that x is a Doppelgänger adversarial to the classifier R iff $\exists y \in \mathfrak{d}(x)$ such that $\text{label}_R(x) \neq \text{label}_R(y)$ and we will refer to both x and y as **adversarial Doppelgängers** when the classifier R is clear from the context.

A classifier R is called **(perceptually) regular** iff it does not admit adversarial Doppelgängers. If R is regular, then $\mathfrak{D}_{\alpha\delta} = \{\mathfrak{d}(x)\}_{x\in\mathbf{X}}$ are R coherent coverings (as defined in [66]).

We say that the **classification problem with m-labels is well defined** if there exists a fully populated and perceptually regular classifier with m labels. Otherwise we say that the classification problem with m-labels is **not well defined**.

The labeling function label_R of a regular classifier is continuous with respect to the perceptual topology τ_δ. Furthermore, if R is not regular and x is **a point of discontinuity** of $\text{label}_R : (\mathbf{X}, \tau_\delta) \to \{1, \ldots, m\}$, then x **is an adversarial Doppelgänger**. The discontinuity is an indication of the cognitive disruption that occurs when one encounters some AD.

It is well known that in some experiences perceptually unambiguous categories and hence perceptually regular classifiers do not exist. A simple example is provided by the perceptual topology in Example 1. Indeed, in this case $\mathbf{X}/\!\sim_\sigma$ is a singleton, i.e, every two inputs are metamorphic and hence there is a only one elementary set which equals the whole $\mathbf{X}$. Therefore, every classifier with two or more labels must have adversarial Doppelgängers.[14]

Clearly, no amount of "robust training" will get rid of adversarial examples of a classifier with a surjective labeling function $\text{label}_R : (\mathbf{X}, \tau_\delta) \to \{1, \ldots, m\}$ if $\mathbf{X}$ cannot be broken into m perceptually unambiguous categories. In the rest of this section we investigate the non existence, existence and internal structure of regular classifiers.

In Part D of the Appendix we show a specific example of a well defined classification problem and discuss the actual regular classifier.

Example 3: If $\overset{\alpha\delta}{\approx}$ is transitive then for every number of labels m smaller than the number of equivalence classes $\text{card}\left(\mathbf{X}/\overset{\alpha\delta}{\approx}\right)$ there exists a fully populated regular classifier with m labels, but if the number of labels m is bigger than $\text{card}\left(\mathbf{X}/\overset{\alpha\delta}{\approx}\right)$ then every fully populated classifier with m labels must admit adversarial Doppelgängers.

Example 3 indicates that if the transitive closure $\sim_\sigma$ is trivial,i.e., $\sim_\sigma = \mathbf{X} \times \mathbf{X}$, then no label is safe. Namely:

Observation 1. *If the transitive closure $\sim_\sigma$ of the indiscriminability relation $\overset{\alpha\delta}{\approx}$ is trivial, then every fully populated classifier with two or more classes admits adversarial Doppelgängers. In particular, let R be a fully populated classifier with a surjective labeling function $\text{label}_R : X \to \{1, 2, \ldots, m\}$, then for every label c there exist adversarial Doppelgängers $x(c) \in \mathbf{X}$ and $x^*(c) \in \mathfrak{d}(x(c))$ such that $c = \text{label}_R\,(x(c))$ and $\text{label}_R\,(x^*(c)) \neq \text{label}_R\,(x(c))$.*

The proof follows from the fact that every finite chain of Doppelgängers connecting points that are labeled differently by a classifier must contain a pair of adversarial Doppelgängers. See Lemma 2 in Appendix Section C.

A straight-forward argument shows that if R is a perceptually regular fully populated classifier, then $x \in R_i$ iff

[10] See Observation 7, in Part B of the Appendix.

[11] A feature $\xi \in \Phi$ is called **attributed** if $\xi \in \Phi_x$ for some input $x \in \mathbf{X}$. It is plausible that $\Phi = \bigcup_{x\in\mathbf{X}} \Phi_x$, and so all features are attributed. However, many models do not preclude the existence of spurious latent traits.

[12] The semantic cluster of inputs sharing the feature $\xi \in \Phi$ is defined for any feature representation. A feature ξ is attributed in a given context, iff $\mathfrak{cl}(\xi) \neq \emptyset$; a feature is a **hypothetical feature**, when $\mathfrak{cl}(\xi) = \emptyset$.

[13] R is fully populated iff $R_c \neq \emptyset$ for every label $c \in \{1, \ldots, m\}$.

[14] See Lemma 3 and the short argument that follows it in Appendix, Section C.

$[x]_{\sim_\sigma} \subset R_i$, i.e., each level set is a disjoint union of elementary sets $[x]_{\sim_\sigma}$. The pigeonhole principle yields:

Observation 2. *If the number of equivalence classes* $\mathrm{card}\,(\mathbf{X}/\sim_\sigma) \geq 2$*, then for every natural number* $2 \leq m \leq \mathrm{card}\,(\mathbf{X}/\sim_\sigma)$ *there exists a perceptually regular fully populated classifier with* m *labels. However, if* $m > \mathrm{card}\,(\mathbf{X}/\sim_\sigma)$*, then every fully populated classifier with* m *labels must have adversarial Doppelgängers.*

In particular, if $p = \mathrm{card}\,(\mathbf{X}/\sim_\sigma) \geq 2$ *is finite, then for every natural number* $m \leq p$ *there are exactly* $S(p, m)$ *regular fully populated classifiers with* m *classes. Here* $S(p, m)$ *is the Sterling number of the second kind.*

Observation 2 shows that the problem of finding a fully populated perceptually unambiguous classifier R with precisely m labels is not well defined if $\mathrm{card}\,(\mathbf{X}/\sim_\sigma) < m$ and vice versa that the same problem is well defined for every m such that $\mathrm{card}\,(\mathbf{X}/\sim_\sigma) \geq m$. In the former case solutions do not exist while in the later case solutions exist and each class segment R_i is a union of equivalence classes $[x]_{\sim_\sigma}$,

$$R_i = \bigcup_{x \in R_i} [x]_{\sim_\sigma}. \tag{5}$$

The existence and properties of discriminative feature representations provide insight whether a classification problem is well defined. In particular, if there exists a discriminative feature representation and the set of attributed features is finite, then every classification problem, whose number of labels exceeds the number of attributed features, is not well defined. See Observation 9 in Appendix Part B.2. Further discussion of the structure of class segments of a regular classifier including the class/category core and fringe are discussed in Section 6.

5. Accuracy and Adversarial Doppelgängers.

The accuracy-adversarial robustness trade off observed and discussed in the literature involves various measures of accuracy [48, 50, 71, 78, 90]. We will discuss classification accuracy. We will show that there is a strong relationship between classifier accuracy and vulnerability to adversarial Doppelgängers. In particular, we will identify (perceptual) scenarios in which low accuracy classifiers are critically vulnerable to adversarial Doppelgänger attacks but on the other hand all high accuracy classifiers can be fooled only by Doppelgängers.

5.1. The Probabilistic Setup.

We will assume that $(\mathbf{X}, \mathcal{F}, \mu)$ is a probability measure space equipped with perceptual topology $\tau_{\mathfrak{d}}$ generated by an indsicriminability relation $\overset{\alpha\mathfrak{d}}{\approx}$ such that for every $x \in \mathbf{X}$ the set of Doppelgängers $\mathfrak{d}(x)$ and the equivalence class $[x]_{\sim_\sigma}$ are events, $(\mathfrak{d}(x) \in \mathcal{F}$ and $[x]_{\sim_\sigma} \in \mathcal{F}, \forall x \in \mathbf{X})$.

In the rest of this section we will assume that the classification problem with $m \geq 2$ labels is well defined and let Ω be a perceptually regular classifier; we will reserve the notation R to denote any classifier (R may or may not be perceptually regular) such that $R_i \cap \Omega_i$, $i = 1, \ldots, m$ are the true positives (of class i). To define accuracy we will focus only on regular models and classifiers s.t., $\Omega_i \in \mathcal{F}$ and $R_i \in \mathcal{F}$, for all $i = 1, \ldots, m$. The **accuracy** of the classifier R defined as

$$\mathrm{accuracy}_\Omega(R) = \mu(R_1 \cap \Omega_1) + \cdots + \mu(R_m \cap \Omega_m). \tag{6}$$

Furthermore let us assume that $\mu(\Omega_i) > 0$ for every $i = 1, \ldots, m$ and thus we can define recall rates

$$\rho_i = \frac{\mu(R_i \cap \Omega_i)}{\mu(\Omega_i)}, \quad i = 1, \ldots, m \tag{7}$$

Bounds on the recall rates imply bounds on the accuracy. Namely if

$$\underline{\rho} \leq \rho_i \leq \bar{\rho}, \quad i = 1, \ldots, m$$

then since μ is a probability measure on X, and

$$\mu(R_i \cap \Omega_i) = \rho_i \mu(\Omega_i), \quad i = 1, \ldots, m \tag{8}$$

we get

$$\underline{\rho} \leq \mathrm{accuracy}_\Omega(R) = \sum_{i=1}^m \rho_i \mu(\Omega_i) \leq \bar{\rho}.$$

5.2. Are Trade-Offs Possible?

Let $i(x) \in \{1, \ldots, m\}$ be the **object class label** of $x \in \mathbf{X}$ i.e., $x \in \Omega_{i(x)}$ and

$$\bar{k}(\Omega) = \sup_{x \in \mathbf{X}} \left(\frac{\mu\left(\Omega_{i(x)}\right)}{\mu(\mathfrak{d}(x))} \right). \tag{9}$$

Every classifier whose recall rates do not exceed $1/\bar{k}(\Omega)$ is totally unsafe in the sense that every correctly classified input admits adversarial Doppelgängers. Specifically:

Observation 3. *Suppose that the sets of Doppelgängers are not negligible and* $\inf_{x \in \mathbf{X}} \mu(\mathfrak{d}(x)) > 0$*, and let* $\Omega = \{\Omega_1, \ldots, \Omega_m\}$ *be a regular world model and let* $R = \{R_1, \ldots, R_m\}$ *be a classifier whose recall rates are strictly smaller than* $1/\bar{k}(\Omega)$ *and so*

$$\frac{\mu\left(R_{i(x)} \cap \mathfrak{d}(x)\right)}{\mu(\mathfrak{d}(x))} \leq \bar{\rho}\bar{k}(\Omega) < 1. \tag{10}$$

Thus every correctly classified input x *has adversarial Doppelgängers.*

Observation 3 shows that sacrificing accuracy may lead to increasing the probability of encountering adversarial Doppelgängers and in fact that there is no trade off for accuracies that are sufficiently low provided that all sets of Doppelgängers have positive measure.

The lack of an opportunity for a trade-off is even more striking when one tries to improve high recall rate (and hence high accuracy) classifiers.

Observation 4. [15] *If* $\inf_{x\in\mathbf{X}} \mu(\mathfrak{d}(x)) > 0$, *and let* $R = \{R_1, \ldots, R_m\}$ *be a classifier whose recall rates are sufficiently high so that* $\underline{\rho} > 1 - 1/\bar{k}(\Omega)$. *i.e.,*

$$\left(1 - \underline{\rho}\right) \bar{k}(\Omega) < 1. \tag{11}$$

Then every misclassified input x *is an adversarial Doppelgänger.*

Definition 6. We say that a classifier has a **hypersensitive behavior** if every misclassified input is an adversarial Doppelgänger.

For classifiers with hypersensitive behavior adversarial robustness can only be improved by improving accuracies, i.e., by eliminating misclassification. Observation 4 shows that if $\mu(\mathfrak{d}(x)) > 0$, for all inputs $x \in \mathbf{X}$, then all classifiers with sufficiently high accuracy are either regular (when accuracy equals to one) or have hypersensitive behavior.

In summary adversarial Doppelgänger robustness - accuracy trade-off may happen for classifiers with middling accuracy rates or when there are inputs whose Doppelgängers are negligible in measure.

6. Life without borders.

A important property of the perceptual topology is that the if a classifier Ω is perceptually regular, then it "imposes an open [topological] borders policy", that is, $\partial\Omega_c = \emptyset$ for every label c. Linguists and psychologists, have observed and postulated that natural perceptual and semantic categories are borderless. See for example [64]. Class/decision boundaries are studied and exploited in many works on classifiers, and in particular on adversarial robustness. These boundaries are artifacts of the metric topology used by the researchers, they are not perceptual phenomena.

However, we are all familiar with the idea that some stimuli are more intrinsic/representative to/of a given class and at the same time frequently there are objects/stimuli/inputs that, while they are firmly with in the class, are less representative/share few(er) features compared to the rest of the elements in the class, that is, they "are/belong to the fringe" of the class.

[15]The proof is in Appendix Part E.

Definition 7. Let s be a similarity scale, i.e., a function $s : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ such that $s(x, x) \geq s(x, y), \forall x, y \in \mathbf{X}$ as in [79], [45] (measuring similarity within a fixed context) and [40].

The values $s(x, x)$ can be and sometimes are used to represent the **salience** or equivalently **importance** of the input within $\mathbf{X}$, see for example [79]. The similarity scale provides a method to quantify the affinity of a input/stimulus to a given measurable subset $D \subset \mathbf{X}$ and the notions of prototype and fringe. The (s-)**affinity** of x with a measurable set D is defined as

$$P\left(x, D\right) = \int_D s(x, y) \tag{12}$$

This is a straightforward generalization of the notion of prtototypicality defined in [79].

Definition 8. $x \in D$ is called a **prototype** of D (with respect to the integrable similarity scale $s : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$) if

$$P\left(x, D\right) = \sup_{z\in D} P\left(z, D\right) \tag{13}$$

$x \in D$ is called a **fringe** element of D (with respect to the integrable similarity scale $s : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$) if

$$P\left(x, D\right) = \inf_{z\in D} P\left(z, D\right) \tag{14}$$

One is tempted to think of prototypes as the stimuli that are "clearest cases, best examples". "easy to tell apart" and to "be a good representative" and hence that optimize salience [13, 31, 63, 64, 79]. However, clearly there is no reason to expect that a stimulus that is unlikely to be observed/encountered would be selected as a prototype. The examples below show that prototypes and fringes are obtained by optimizing a mixture of the frequency of appearance (likelihood to encounter) and salience.

The core and fringe sets may be empty. In practice one is often satisfied with finding elements that may not be true prototypes but belong to the M**-core**, i.e., have affinity exceeding a fixed threshold M. Similarly, elements whose affinity is below a given threshold τ belong to the τ**-fringe**.

Many perceptual and cognitive processes exploit the intelligent agents' abilities to measure and/or compare the salience/importance of features. In some simple cases it is expected and it even might be true that salience/importance is a probability measure f_Φ on on the space of (all possible) features Φ and the feature representations Φ_x are measurable subsets of Φ. In many accounts including Tversky's feature contrast model [79] the **salience scale** f_Φ is a context dependent, nonnegative function defined on a collection $\Upsilon(\Phi)$ of subsets of Φ which is closed under finite unions and intersections, and set differences. Furthermore, $\Phi_x \in \Upsilon(\Phi), \forall x \in \mathbf{X}$; and the non-negative function f_Φ is **feature additive**, i.e, $f_\Phi(A \cup B) = f_\Phi(A) +$

$f_\Phi(B)$, if $A \cap B = \emptyset$. The value $f(x) = f_\phi(\Phi_x), \forall x \in \mathbf{X}$ is called the **salience/prominence of the input** x, [79]. We will call a (Tversky) salience scale f_Φ perceptually regular if the prominence/salience function $f(x) = f_\phi(\Phi_x)$ is perceptually regular. We will call a perceptually regular salience scale f_Φ **fully deployable** if it can be used to judge the distinguishing features of inputs/stimuli. In particular, we can use $f_\Phi(\Phi_x \setminus \Phi_y)$ to discriminate x from y. Thus if f_Φ is fully deployable, then $f_\Phi(\Phi_x \setminus \Phi_y) = 0$ and $f_\Phi(\Phi_y \setminus \Phi_x) = 0$ whenever $x \overset{\alpha\delta}{\approx} y$ or equivalently $f_\Phi(\Phi_x \cap \Phi_y) = f(x)$ if $x \overset{\alpha\delta}{\approx} y$.

The following special case provides a particularly useful insight into the nature of prototypes and the possible internal structure of categories as discussed in [64].

Example 4: If s is a contrast similarity [79] such that $f_\Phi(\Phi_x) = s(x,x)/\theta$ is fully deployable, $\theta \in (0, +\infty)$, and furthermore, $f_\Phi(\Phi_x \cap \Phi_y) = f_\Phi(\Phi_x)$ for every pair $x \sim_\sigma y$ and $\Phi_x \cap \Phi_y = \emptyset$ if $x \not\sim_\sigma y$, then for every regular class $D \subset \mathbf{X}$, and every $x \in D$ we get

$$P(x, D) = \Theta\, \mu(D) \left(\frac{\mu([x]_{\sim_\sigma})}{\mu(D)} - \frac{\alpha}{\Theta} \right) f_\Phi(\Phi_x) - \beta I_\Phi(D), \tag{15}$$

where $\Theta = (\alpha + \beta + \theta)$ and α, β, and θ are non-negative constants. In particular, if D is a finite union of equivalence classes $\zeta_j \in \mathbf{X}/\sim_\sigma$, then both prototypes and fringe elements exist. Furthermore, if x is prototype/fringe then so are all stimuli in its component $\zeta = [x]_{\sim_\sigma}$. Similar statements hold for M-core and τ-fringe elements.

If the class D is fixed, then $\dfrac{\mu\left([x]_{\sim_\sigma}\right)}{\mu(D)}$ is just the probability to encounter (and learn) a set of stimuli, and $f_\Phi(\Phi_x)$ can be interpreted as the level of prominence. So the optimization process involves learning prominent examples that can be encountered reasonably often. Thus in this case the prototype does not fall in either of the two main branches of prototypes, that is inputs that represent the central tendency in the regular class vs. prototypes as highly "representative exemplar(s) of a category", see [15], page 52.

More generally, when the similarity scale is bounded, for example, this is true for real similarity measures deployable by humans, then, as predicted by modern prototype theory, each perceptually regular subset $D \subset \mathbf{X}$ corresponding to real (natural) categories created and analyzed by real intelligent agents consists of core elements (possibly M-core), a layer of fringe (possibly τ-fringe) elements, and layers of elements of various levels of intermediate affinity with D. In Part I of the Appendix we introduce class invariants of regular classifiers including structural entropy, expected index of coincidence, and importance.

7. Quantifying Doppelgänger Vulnerability.

By definition if a classifier R is not regular, then it is vulnerable to adversarial Doppelgängers attacks, that is, for some input x there exists $a(x) \overset{\alpha\delta}{\approx} x$ and such that $\text{label}_R(x) \neq \text{label}_R(a(x))$. In particular, we say that R is **conceptually ambiguous** at x and we call the set

$$\mathrm{A}(R) = \{x \in \mathbf{X} : \exists y \overset{\alpha\delta}{\approx} x \text{ and } \text{label}_R(x) \neq \text{label}_R(y)\} \tag{16}$$

the **region of conceptual ambiguity**. When (X, μ) is a probability measure space and $\mu(\mathfrak{d}(x)) > 0$, we use the probability distribution of labels at x:

$$p_j(x) = \frac{\mu\left(R_j \bigcap \mathfrak{d}(x)\right)}{\mu(\mathfrak{d}(x))}, \quad j = 1, \ldots, m \tag{17}$$

and the **conceptual entropy** of R at x defined as

$$H_R(x) = -\sum_{j=1}^{m} p_j(x) \log(p_j(x)) \tag{18}$$

to detect whether R is conceptually ambiguous at x (i.e., $H_R(x) > 0$), and to quantify the likelihoods of various adversarial Doppelgänger attacks.

Definition 9. Let R be a classifier and $\hat{a} : \mathbf{X} \to \mathbf{X}$, we will call the inner measure of the set $\{x : \text{label}_R(\hat{a}(x)) \neq \text{label}_R(x)\}$ the R-**fooling rate** of the mapping $\hat{a}$, and we will denote it by $F_R(\hat{a})$. A mapping $\hat{a} : \mathbf{X} \to \mathbf{X}$ is called an **adversarial Doppelgänger attack** to a classifier R if and only if $\hat{a}(x) \overset{\alpha\delta}{\approx} x$, $\forall x \in \mathbf{X}$, and the R-fooling rate $F_R(\hat{a})$ is positive.

The set $\{x : \text{label}_R(\hat{a}(x)) \neq \text{label}_R(x)\}$ is a subset of the region of conceptual ambiguity of R, which yields an upper bound on the R-fooling rate by the outer measure of $A(R)$:

$$F_R(\hat{a}) \leq \mu_*(A(R)). \tag{19}$$

In specific scenarios it is possible to get an upper bound on the size, possibly the outer measure, of $A(R)$ which in turn shows that the R-fooling rates are bounded away from one. See Example 11 in Part J in the Appendix.

Adversarial Doppelgänger attacks are distinct from the adversarial attacks studied to date. The universal adversarial attacks, [47], can achieve fooling rates as close to one as one desires. As illustrated above, adversarial Doppelgänger attacks may not be able to reach fooling rates that are too high. On the other hand in some cases, the optimal fooling rate of one can be achieved.

Observation 5. *If R is conceptually ambiguous at every $x \in \mathbf{X}$, e.g., when $H_R(x) > 0$ for every $x \in \mathbf{X}$, then there exists an adversarial Doppelgänger attack with R-fooling rate equal to one.*

Indeed, if R is conceptually ambiguous at every $x \in \mathbf{X}$, then $\{y \in \mathfrak{d}(x) : \text{label}_R(y) \neq \text{label}_R(x)\} \neq \emptyset$, for every x, and therefore the axiom of choice implies that there exists a map $\hat{a} : \mathbf{X} \to \mathbf{X}$ such that $\hat{a}(x) \in \{y \in \mathfrak{d}(x) : \text{label}_R(y) \neq \text{label}_R(x)\}$, for every $x \in X$. It turns out that in practice there may be many classifiers that are conceptually ambiguous at every input.

Example 5: Consider the case when $\overset{\alpha\delta}{\approx}$ is transitive, i.e., $\overset{\alpha\delta}{\approx}$ equals its transitive closure $\sim_\sigma$, e.g, when $\mathbf{X}$ is equipped with a Grassmann structure as in [39], and there exist at least two different equivalence classes. Thus binary classification is a well defined problem. There exist at least $\prod_{\zeta \in \mathbf{X}/\sim_\sigma} \{U \subset \zeta : 0 < \mu(U) < \mu(\zeta)\}$ worth of fully-populated binary classifiers which are conceptually ambiguous at every input $x \in \mathbf{X}$.

The last example and Observation 5 show that even high accuracy classifiers can be vulnerable to adversarial Doppelgänger attacks with fooling rate equal to one.

We conclude this section with a warning that popular methods to deal with unseen data, including **marking missing data and imputation, may introduce conceptual ambiguity**. For example, if a model is trained on a data set $T \subset \mathbf{X}$ that includes only parts of some elementary sets, then adding a class label NA to label unseen data can compromise adversarial Doppelgänger robustness. Indeed, let S be a nonempty set of training data such that $S \subsetneq \zeta \in \mathbf{X}/\sim_\sigma$. There exists, $x \in S$ such that $\mathfrak{d}(x) \setminus S \neq \emptyset$. Every $z \in \mathfrak{d}(x) \setminus S$, labeled as NA, is an adversarial Doppelgänger of x.

8. Discussion and Conclusions.

A central focus of this paper is the adversarial Doppelgängers phenomenon, where classifiers assign different labels to inputs that humans cannot discriminate. Until now, this phenomenon has not been well understood, possibly due to the limitations of the distance-based analysis that has dominated the field. In the "absence of a distance measure that accurately captures the perceptual differences between a source and adversarial example many researchers have decided to use the ℓ_p distance", [29]. The available empirical observations and models - both perceptual and cognitive, including those based on just noticeable differences - provide no evidence that biologically plausible perceptual topologies are metric. This paper advances the understanding of context-related perceptual topologies in input spaces, which are rarely metric. Our investigation shows that adversarial Doppelgängers are very close to each other with respect to the context-relevant perceptual metric d_w, this metric is not a manifold metric and does not generate the perceptual topology. This distinction highlights the shortcomings of traditional, purely manifold metric-based representations and analysis of perceptual spaces.

The machine learning community has expended significant efforts aimed to build adversarially robust classifiers. This may be a march towards a bridge too far. Philosophers, experimental psychologists, and linguists, are well aware that many classification problems are not well defined due to perceptual ambiguities. Any fully populated classifier for a classification problem that is not well defined is doomed to be a victim of the adversarial Doppelgängers phenomenon. Our results reveal the structure of adversarial Doppelgänger-robust classifiers, regular classifiers, and criteria and methods to establish whether a classification problem is well defined or not. The new understanding of the structure of regular classifiers, the analysis of zones of ambiguity, and the methods to measure and bound the fooling rates of adversarial Doppelgänger attacks provide guidance on how to design adversarially robust training to improve classifiers that are not regular. In addition to revealing the impossibility to use accuracy-robustness trade-offs in many scenarios, including robustifying hypersensitive classifiers, our analysis indicates that marking unseen data can jeopardize robustness if the training data contains only a proper subset of an elementary set.

We explore feature representations, the related concept of indiscernibility introduced by Leibniz, and their connection to indiscriminability. This investigation reveals the nature of class prototypes and fringe inputs, and how the size of a discriminative feature representation can be used to determine whether a classification problem is not well defined. Indiscernibility and indiscriminability, are often conflated in the machine learning literature. Elucidating the distinction between them is vital for understanding the limitations of current classifiers and addressing the shortcomings in their design.

Our discussion of the Doppelgängers phenomenon brings to light a significant divergence between human perception and artificial neural network models, including feedforward models, RNN models and ResNet. The indiscriminability relations of these artificial neural network models, studied in [18], are transitive[16] while it is well accepted, that, in many contexts, the human indiscriminability relation is not transitive.

The results and insights gained from this investigation point to concrete warnings and actionable steps for improving the training and testing of classifiers.

[16]See Part K in the Appendix.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 977–988. IEEE, 2022. 2

[2] Valtteri Arstila. Why the transitivity of perceptual simultaneity should be taken seriously. *Frontiers in integrative neuroscience*, 6:3, 2012. 3

[3] Benjamin Balas, Lisa Nakano, and Ruth Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of vision*, 9(12):13–13, 2009. 2, 3

[4] Max Black. The identity of indiscernibles. *Mind*, 61(242): 153–164, 1952. 1

[5] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1987–2004. IEEE, 2022. 2

[6] William F Broderick, Gizem Rufo, Jonathan Winawer, and Eero P Simoncelli. Foveated metamers of the early visual system. *bioRxiv*, pages 2023–05, 2023. 1, 2, 3, 6

[7] Fan RK Chung. Lectures on spectral graph theory. *CBMS Lectures, Fresno*, 6(92):17–21, 1996. 6

[8] Samuel Clarke. *A Collection of Papers, which Passed Between the Late Learned Mr. Leibnitz, and Dr. Clarke, in the Years 1715 and 1716: Relating to the Principles of Natural Philosophy and Religion. With an Appendix. To which are Added, Letters to Dr. Clarke Concerning Liberty and Necessity; from a Gentleman of the University of Cambridge: with the Doctor's Answers to Them. Also Remarks Upon a Book, Entituled, A Philosophical Enquiry Concerning Human Liberty*. James Knapton, at the Crown in St. Paul's Church-Yard., 1717. 3

[9] Jozef B Cohen and William E Kappauf. Color mixture and fundamental metamers: Theory, algebra, geometry, application. *The American journal of psychology*, pages 171–259, 1985. 2, 3

[10] Rafael De Clercq and Leon Horsten. Perceptual indiscriminability: In defence of Wright's proof. *The Philosophical Quarterly*, 54(216):439–444, 2004. 3

[11] Lieven Decock and Igor Douven. Similarity after Goodman. *Review of philosophy and psychology*, 2:61–75, 2011. 2

[12] Kaize Ding, Albert Jiongqian Liang, Bryan Perrozi, Ting Chen, Ruoxi Wang, Lichan Hong, Ed H. Chi, Huan Liu, and Derek Zhiyuan Cheng. Hyperformer: Learning expressive sparse feature representations via hypergraph transformer, 2023. 2

[13] Igor Douven. Putting prototypes in place. *Cognition*, 193: 104007, 2019. 6

[14] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010. 2

[15] William K Estes. *Classification and cognition*. Oxford University Press, 1994. 7

[16] Katalin Farkas. *The subject's point of view*. OUP Oxford, 2010. 2, 3

[17] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *Advances in neural information processing systems*, 31, 2018. 2

[18] Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 8, 7

[19] Gustav Theodor Fechner. *Elemente der psychophysik*. Breitkopf u. Härtel, 1860. 3

[20] Gustav Theodor Fechner. *Revision der hauptpuncte der Psychophysik*. Breitkopf und Härtel, 1882. 3

[21] Peter Forrest. The Identity of Indiscernibles. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020. 3

[22] Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011. 2, 3

[23] Peter Gardenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2004. 2

[24] Amir Globerson and Sam Roweis. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on Machine learning*, pages 353–360, 2006. 2

[25] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[26] Nelson Goodman. *The structure of appearance*. Harvard Univ. Press, 1951. 3

[27] Delia Graff. Phenomenal continua and the sorites. *Mind*, 110 (440):905–936, 2001. 3

[28] Hermann Grassmann. Zur theorie der farbenmischung. *Annalen der Physik*, 165(5):69–84, 1853. 3

[29] Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 43–49. IEEE, 2018. 8

[30] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020. 2

[31] Eleanor R Heider. " Focal" color areas and the development of color names. *Developmental psychology*, 4(3):447, 1971. 6

[32] Benj Hellie. Noise and perceptual indiscriminability. *Mind*, 114(455):481–508, 2005. 3

[33] Christopher James Henry. *Near sets: theory and applications*. University of Manitoba (Canada), 2011. 2

[34] Felix Hovsepian. *A metalogical analysis of vagueness: an exploratory study into the geometry of logic*. PhD thesis, University of Warwick, 1992. 1

[35] Ke Huang and Selin Aviyente. Sparse representation for signal classification. *Advances in neural information processing systems*, 19, 2006. 2

[36] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 2

[37] Akshay V Jagadeesh and Justin L Gardner. Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, 119(17):e2115302119, 2022. 1, 2, 3

[38] Ralph Kopperman. All topologies come from generalized metrics. *The American Mathematical Monthly*, 95(2):89–97, 1988. 1

[39] David H Krantz. Color measurement and color theory: I. representation theorem for grassmann structures. *Journal of Mathematical Psychology*, 12(3):283–303, 1975. 2, 3, 8

[40] Dekang Lin et al. An information-theoretic definition of similarity. In *ICML*, pages 296–304, 1998. 6

[41] Alexander D Logvinenko. Object-colour manifold. *International journal of computer vision*, 101:143–160, 2013. 2, 3

[42] Alexander D. Logvinenko and Eugene Demidenko. On counting metamers. *IEEE Transactions on Image Processing*, 25(2):770–775, 2016. 2, 3

[43] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[44] Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4536–4543, 2019. 2

[45] Douglas L Medin, Robert L Goldstone, and Dedre Gentner. Respects for similarity. *Psychological review*, 100(2):254, 1993. 2, 6

[46] Alexius Meinong. *Über die Bedeutung des Weber'schen Gesetzes: Beiträge zur Psychologie des Vergleichens und Messens*. L. Voss, 1896. 3

[47] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 2, 7

[48] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 5

[49] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997. 2

[50] Marwan Omar, Soohyeon Choi, DaeHun Nyang, and David Mohaisen. Robust natural language processing: Recent advances, challenges, and future directions. *IEEE Access*, 2022. 5

[51] Wilhelm Ostwald. *Physikalische Farbenlehre*. Verlag Unesma, 1919. 2, 3

[52] 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)*, pages 372–387. IEEE, 2016. 2, 5

[53] Zdislaw Pawlak. *Classification of objects by means of attributes*. Polish Academy of Sciences [PAS]. Institute of Computer Science, 1981. 3

[54] Charles Pelling. Exactness, inexactness, and the nontransitivity of perceptual indiscriminability. *Synthese*, 164 (2):289–312, 2008. 3

[55] Mika Perälä. Aristotle on perceptual discrimination. *Phronesis*, 63(3):257–292, 2018. 1

[56] James F Peters. Near sets. Special theory about nearness of objects. *Fundamenta Informaticae*, 75(1-4):407–433, 2007. 2

[57] James F Peters and Piotr Wasilewski. Tolerance spaces: Origins, theoretical aspects and applications. *Information Sciences*, 195:211–225, 2012. 1, 2

[58] Henri Poincaré. *Dernières pensées*. Flammarion, 1930. 2

[59] Tim Poston. *Fuzzy geometry*. PhD thesis, University of Warwick, 1971. 2, 3, 1

[60] Diana Raffman. Is perceptual indiscriminability nontransitive? *Philosophical Topics*, 28(1):153–175, 2000. 2

[61] Diana Raffman. Indiscriminability and phenomenal continua. *Philosophical perspectives*, 26:309–322, 2012. 3

[62] Fred S Roberts. Tolerance geometry. *Notre Dame Journal of Formal Logic*, 14(1):68–76, 1973. 2

[63] Eleanor Rosch. Cognitive reference points. *Cognitive Psychology*, 7(4):532–547, 1975. 6

[64] Eleanor H Rosch. On the internal structure of perceptual and semantic categories. In *Cognitive development and acquisition of language*, pages 111–144. Elsevier, 1973. 6, 7

[65] Bertrand Russell. *The principles of mathematics*. Routledge, 2020. 3

[66] Julius A Schreider. *Equality, Resemblance and Order*. Mir Publishers, Moscow, 1975. 2, 4

[67] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018. 2

[68] Herbert A Simon. *Models of man; social and rational.* Wiley, 1957. 3

[69] Alexei B Sossinsky. Tolerance space theory and some applications. *Acta Applicandae Mathematica*, 5:137–167, 1986. 6

[70] Michael W Spratling. Classification using sparse representations: a biologically plausible approach. *Biological cybernetics*, 108:61–73, 2014. 2

[71] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?– a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 2, 5

[72] Jeremias Sulam, Ramchandran Muthukumar, and Raman Arora. Adversarial robustness of supervised sparse coding. *Advances in Neural Information Processing Systems*, 33:2110–2121, 2020. 2

[73] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[74] Thomas Tanay and Lewis Griffin. A boundary tilting perspective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016. 2

[75] Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4 (Dec):1235–1260, 2003. 2

[76] William A Thornton. Matching lights, metamers and human visual response. *The Journal of Color and Appearance*, 2: 23–29, 1973. 2, 3

[77] Edward Chace Tolman. *Purposive behavior in animals and men*. Univ of California Press, 1932. 2

[78] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 2, 5

[79] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977. 2, 6, 7

[80] Gottfried Wilhelm Freiherr von Leibniz. *Leibniz: Discourse on Metaphysics*. Open Court Publishing Company, 1902. 3, 2

[81] Stella Vosniadou and Andrew Ortony. *Similarity and analogical reasoning*. Cambridge University Press, 1989. 2

[82] Ernst Heinrich Weber. *De Pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae...* CF Koehler, 1831. 3

[83] Ernst Heinrich Weber. Tastsinn und gemeingefuhl. *Handworterbuch der Physiologie*, 1846. 3

[84] Timothy Williamson. *Identity and Discrimination*. Wiley-Blackwell, 1990. 2, 3

[85] Crispin Wright. On the coherence of vague predicates. *Synthese*, pages 325–365, 1975. 3

[86] Günter Wyszecki. Valenzmetrische untersuchung des zusammenhanges zwischen normaler und anomaler trichromasie. 1953. 2, 3

[87] Günther Wyszecki and Walter Stanley Stiles. *Color science: concepts and methods, quantitative data and formulae*. John Wiley & Sons, 2000. 3

[88] Erik C. Zeeman. The topology of the brain and visual perception. In *The Topology of 3-Manifolds*. Prentice Hall, NJ, 1962. 2

[89] E. Christopher Zeeman. Topology of the brain. In *Math and Computer Science in Biology and Medicine*. H.M. Stationary Office, London, United Kingdom, 1965. 2

[90] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 2, 5

[91] Zheng Zhang, Yong Xu, Jian Yang, Xuelong Li, and David Zhang. A survey of sparse representation: algorithms and applications. *IEEE access*, 3:490–530, 2015. 2

[92] Li Zhaoping and Keith A May. Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex. *PLoS Computational Biology*, 3(4):e62, 2007. 3