bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Attack and defence in cellular decision-making: lessons from machine learning

Thomas J. Rademaker[1], **Emmanuel Bengio**[2], **Paul François**[1]

***For correspondence:** paul.francois2@mcgill.ca (Paul François)

[1] Ernest Rutherford Physics Building, McGill University, 3600 rue University, H3A2T8 Montreal, QC, Canada; [2] School of Computer Science, McGill University, 3480 rue University, H3A0E9 Montreal, QC, Canada

Abstract Machine learning algorithms are sensitive to meaningless (or "adversarial") perturbations. This is reminiscent of cellular decision-making where ligands (called "antagonists") prevent correct signalling, like in early immune recognition. We draw a formal analogy between neural networks used in machine learning and models of cellular decision-making (adaptive proofreading). We apply attacks from machine learning to simple decision-making models, and show explicitly the correspondence to antagonism by weakly bound ligands. Such antagonism is absent in more nonlinear models, which inspired us to implement a biomimetic defence in neural networks filtering out adversarial perturbations. We then apply a gradient-descent approach from machine learning to different cellular decision-making models, and we reveal the existence of two regimes characterized by the presence or absence of a critical point. The critical point causes the strongest antagonists to lie close to the threshold. This is validated in the loss landscapes of robust neural networks and cellular decision-making models, and observed experimentally for immune cells. For both regimes, we explain how associated defence mechanisms shape the geometry of the loss landscape, and why different adversarial attacks are effective in different regimes. Our work connects evolved cellular decision-making to machine learning, and motivates the design of a general theory of adversarial perturbations, both for *in vivo* and *in silico* systems.

Introduction

Machine learning is becoming increasingly popular with major advances coming from deep neural networks [31]. Deep learning has improved the state-of-the-art in automated tasks like image processing [24], speech recognition [19] and machine translation [44], and has already seen a wide range of applications in research and industry. Despite their success, neural networks suffer from blind spots: small perturbations added to unambiguous samples may lead to misclassification [45]. Such adversarial examples are most obvious in image recognition, for example, a panda is misclassified as a gibbon or a handwritten 3 as a 7 [16]. Real world scenarios exist, like adversarial road signs fooling computer vision algorithms (Fig. 1 A) [39]. Worse, adversarial examples are often transferable across algorithms (see [1] for a recent review), and certain "universal" perturbations fool any algorithm.

Categorization and inference are also tasks found in cellular decision-making. For instance, T cells have to discriminate between foreign and self ligands which is challenging since foreign ligands might not be very different biochemically from self ligands [14, 10]. Decision-making in an immune context is equally prone to detrimental perturbations in a phenomenon called ligand antagonism [11]. Antagonism appears to be a general feature of cellular decision-makers: it has been observed in T cells [2], mast cells [47] and other recognition processes like olfactory sensing [41].

There is a natural analogy to draw between decision-making in machine learning and in biology. In machine learning terms, cellular decision-making is similar to a classifier. Furthermore, in both artificial and cellular decision-making, targeted perturbations lead to faulty decisions even in the presence of a clear "ground truth" signal. As a consequence, arms races are observed in both systems. Mutating agents might systematically explore ways to fool the immune cells via antagonism, as has been proposed in the HIV case [23, 35, 21]. This is reminiscent of how adversaries could generate

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

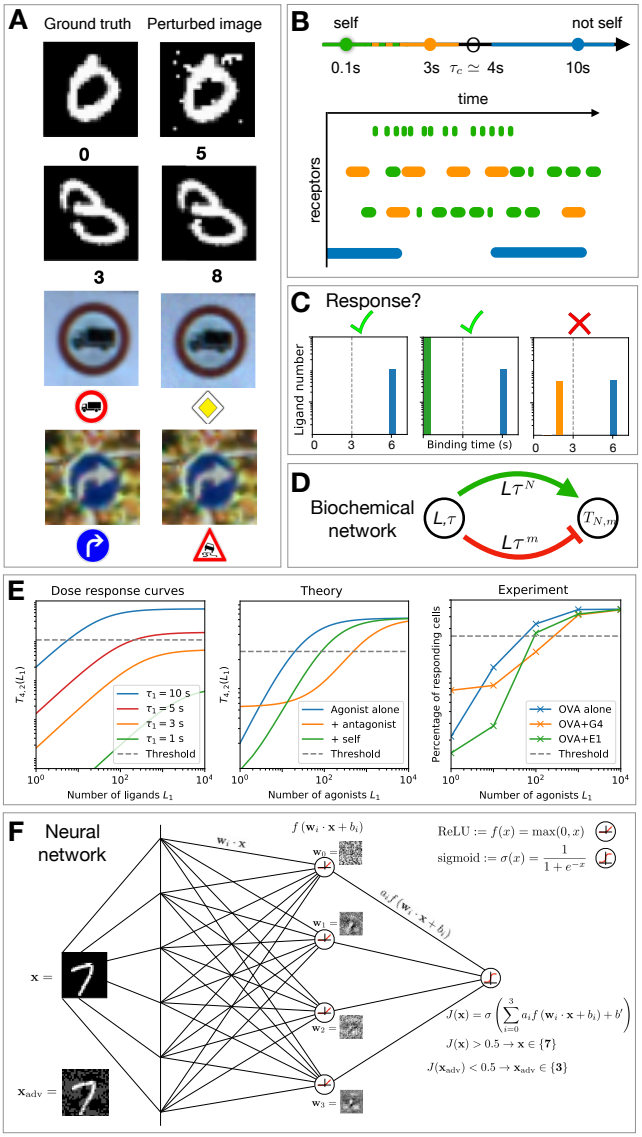


Figure 1. Ligand discrimination and digit recognition. A) Adversarial examples on digits and roadsigns. Reproduced from [39]. Left column displays original images with categories recognized by machine learning algoritms, right column displays targeted perturbations of images leading to incorrect classifications B) Schematics of ligand binding events showing typical receptor occupancy through time for cellular decision making, using a T cell terminology ("self vs non self") C) Different ligand distributions give different response. Vertical dotted line indicate quality τ_c, decision should be taken if one observes ligands with $\tau > \tau_c$, so on the right of this line. In a T cell context, cells responds to ligand distributions of agonists alone and agonists in the presence of self (with very small binding times τ), while the T cell fails to respond if there are too many ligands just below threshold τ_c. D) Schematic of adaptive proofreading networks, with both activating and repressing branches, with different weights of τ. E) Dose-response curves for ligands with different binding times for pure ligand types and for mixtures, in both adaptive proofreading models and experiments on T cells (redrawn from [13]). The offset in the denominator is set to 3000, $\tau_c = 4s$, mixtures consist of L_1 ligands at $\tau_{\text{ag}} = 10s$ and $L_2 = 10^4$ ligands at $\tau_a = 3s$ or $\tau_{\text{self}} = 1s$. For experiments, OVA are agonist ligands, G4 and E1 are ligands known to be below threshold, but showing clear antagonistic properties. F) Schematic of the neural network used for digit recognition. We explicitly show the 4 matrices W_i learnt in one instance of the training, final activation function J and one example of adversarially perturbed sample $\mathbf{x}_{adv}$

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

"black box attacks" aiming at fooling neural networks [39]. Strategies for provable defenses or robust detection of adversarial examples [18, 51] are currently developed in machine learning, but we are still far from a general solution.

Adaptive proofreading for cellular decision-making

Cellular decision-making in our context refers to classification of biological ligands in two categories, e.g. "self vs non self" in immunology, or "agonist vs non agonist" in endocrinology. For most of those cases, there actually is a continuum of properties between two different categories, so that it is convenient to rank different ligands based on a continuous variable (notation τ) that we will call "quality". Mathematically, a cell needs to decide if it is exposed to ligands with quality $\tau > \tau_c$, where τ_c is the quality at the decision threshold. Such ligands triggering response are called "agonists". A general problem then is to consider cellular decision-making based on ligand quality irrespective of ligand quantity (notation L). An example can be found in immune recognition with the "lifetime dogma" [10], where it is assumed that a T cell discriminates ligands based on their characteristic binding time τ to T cell receptors (this is of course an approximation and other parameters might also play a role in defining quality, see [17, 6, 33]). Ligand discrimination is a nontrivial problem for the cell, which does not measure single-binding events but only has access to global quantities such as the total number of bound receptors (Fig. 1 B). The challenge is to ignore many subthreshold ligands ($\tau < \tau_c$) while responding to few agonist ligands with $\tau > \tau_c$ [2, 10, 13]. In particular, it is known experimentally in many different contexts that addition of "antagonistic" subthreshold ligands can impair proper decision-making (Fig. 1 C) [2, 47, 41].

To model cellular decision-making, we will use the general class of "adaptive sorting" or "adaptive proofreading", models, which account for many aspects of immune recognition [29, 11], and can be shown to capture all relevant features of such cellular decision-making close to a decision threshold [12]. Mathematical details on the models are given in Appendix 1. Adaptive proofreading models rely on an incoherent feedforward loop, where an output is at the same time activated and repressed by bound ligands via two different branches in a biochemical network (Fig. 1 D). Thanks to the tug-of-war between those different branches, the network can perform a complex computation. The activation and repression branch are assumed to be activated linearly as a function of the ligand quantity L. We further assume that the τ dependency of the activation branch is stronger (parameter N, generally assumed to be an integer representing number of phosphorylation sites in immune recognition models) than the repression branch (parameter m, we assume $m < N$). The output $T_{N,m}$ (taken to be the ratio between the activation and the repression branch) ensures that the ligand dependency L disappears while a τ dependency survives, allowing for ligand classification based on quality τ.

Fig. 1 E shows theoretical and experimental curves of an adaptive proofreading model with $(N, m) = (4, 2)$. Adaptive proofreading models give dose response curves plateauing at different values as a function of parameter τ, allowing to perform sensitive and specific measurement of this parameter (Fig. 1 E left). For small τ (e.g. $\tau = 3$s), one never reaches the detection threshold (dotted line on Fig. 1 E) even for many ligands. For slightly bigger $\tau = 10 \quad \mathrm{s} > \tau_c$, the curve is shifted up so that detection is made even for a small concentration of agonists.

If we now consider mixtures of ligands with different qualities, the respective computation made by the activation and repression branch of the network depends in different ways on τs. Antagonistic effects occur when the repression branch is activated more strongly than the activation branch, thus killing the response. This corresponds to the "dog in the manger" effect described in [47] for decision-making by mast cells. Fig. 1 E middle and right panels illustrate this. In presence of many ligands below the threshold of detection, the dose response curve for "agonist alone" are simultaneously moved to the right but with a higher starting point, as observed experimentally (data redrawn from [13]). Different models have different antagonistic properties, based on the strength of the activation branch (N) relative to the repression branch (m). More mathematical details on such models can be found in [29, 11, 12].

Neural networks for artificial decision-making

We will compare cellular decision-making to decision-making in machine learning algorithms. We will constrain our analysis to binary decision-making, here classifying images from two types of digits.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

These images are taken from MNIST [32], a standard database with 70000 pictures of handwritten digits. Even for such a simple task, designing a good classifier is not trivial, since it should be able to classify irrespective of subtle changes in shapes, intensity and writing style (i.e. with or without a central bar for a 7).

A simple machine learning algorithm is logistic regression. Here, the inner product of the input and an optimized weight matrix determines the class of the input. Another class of machine learning algorithms are feedforward neural networks: interconnected groups of nodes processing information layer-wise. We chose to work with neural networks for several reasons. First, logistic regression is a limiting case of a neural network without hidden layers. Second, a neural network with at least one hidden layer more closely imitates information processing in cellular networks. Third, such an architecture reproduces classical results on adversarial perturbations such as the ones described in [16]. Fig. 1 E introduces the iterative matrix multiplication inside a neural network. Each neuron i computes $\mathbf{w}_i \cdot \mathbf{x}$, $i \in [0, 3]$, adds bias b_i, and transforms the result with an activation function $f(x)$. We chose to use a Rectified Linear Unit (ReLU), which returns 0 when its input is negative, and the input itself otherwise. The resulting $f(\mathbf{w}_i \cdot \mathbf{x} + b_i)$ is multiplied by another weight matrix with elements a_i, summed up with a bias, defining a scalar quantity $x = \sum_i a_i f(\mathbf{w}_i \cdot \mathbf{x} + b_i) + b'$. Finally, the logistic function $\sigma(x)$ assigns a class (0 or 1) probability to the input.

We use the scikitlearn implementation [40] of the multilayer perceptron to train neural networks. We have chosen our hyperparameters as follows: one hidden layer with four neurons feeding into an output neuron, a random 80/20 training/test split with a 10 percent validation split. The cross-entropy loss function is minimized via stochastic gradient descent in maximal 300 iterations with a batch size of 200 and an adaptive learning rate, initiated at 0.001. The tolerance is 10^{-4} and the regularization rate is 0.1. Most of these parameters are set to their default value, but we found that the training procedure is largely insensitive to the specific choice of hyperparameters. As an example, in Fig. 1 E, a 7 is correctly classified by the neural network ($J(\mathbf{x}) > 0.5$), while the adversarial 7 is classified as a three ($J(\mathbf{x}_{\text{adv}}) < 0.5$).

Results

We first summarize the general approach followed to draw the parallel between machine learning and cellular decision-making. We will limit ourselves to simple classifications where a single decision is made, such as "agonist present vs no agonist present" in biology, or "3 vs 7" in digit recognition. Decision-making on a sample (a picture in machine learning, or a ligand distribution in biology) is then done via a scoring function (or score). This score is computed either directly by the machine learning algorithm (score J) or by the biochemical network, via the concentration of a given species ($T_{N,m}$ where (N, m) depend on the model considered, see Appendix 1). For simple classifications, the decision is then based on the relative value of the score above or below some threshold (typically 0.5 for neural networks where decision is based on sigmoidal functions, or some fixed value related to critical quality τ_c for biochemical networks).

The overall performance of a given classifier depends on the behavior of the score in the "space" of possible samples (i.e. the space of all possible pictures, or the space of all possible ligand distributions). Both spaces have extremely high dimension: for instance dimension in the MNIST picture correspond to number of pixels $28 \times 28 = 784$, while in immunology there are roughly 30000 receptors potentially bound to different ligands [2]. The score can thus be thought of as a non-linear projection in one dimension of samples in those huge spaces. We will study how the score behaves in relevant directions in the sample space, and how to change the corresponding geometry and position of decision boundaries (defined as the samples where the score is equal to the classification threshold). We will show that similar properties are observed, both close to initial samples and to the decision boundary. It is important to notice at this stage that the above considerations are completely generic on the biology side and are not necessary limited to, say immune recognition; however we will show that adaptive proofreading presents many features reminiscent of what is observed in machine learning.

Fast Gradient Sign Method recovers antagonism by weakly binding ligands

In this framework, from a given sample, an adversarial perturbation is a "small" perturbation in sample space giving a change in score reaching (or crossing) the decision boundary. We start by mathematically

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

connecting the simplest class of adversarial examples in machine learning to antagonism in adaptive proofreading models. We follow the original Fast Gradient Sign Method (FGSM) proposed by [16]. The FGSM computes the local maximum adversarial perturbation $\eta = \epsilon \mathrm{sgn}\left(\nabla_x J\right)$ with $||\eta||_\infty \leq \epsilon$. $\nabla_x J$ represents the gradient of the scoring function, categorizing images in two different categories (such as 3 and 7 in [16]). Its sign defines an image, that is added to the initial batch of images with small weight ϵ. Examples of such perturbations are shown in Fig. 1 E and Appendix 2 for the 3 vs 7 digit classification problem. While to the human observer, the perturbation is weak and only changes the background, "naive" machine learning algorithms are completely fooled by the perturbation and systematically misclassify the digit.

Coming back to adaptive proofreading models, we apply FGSM for the computation of a maximally antagonistic perturbation. To do so, we need to specify the equivalent of pixels in adaptive proofreading models. A natural choice is to consider parameters associated to each pair (index i) of receptor/ligands, namely k_i (corresponding to an on-rate) and τ_i (corresponding to quality), and to compute gradients with respect to those parameters. If a receptor i is unoccupied, we make the choice to consider that this k_i and τ_i are 0 [1].

As a simple example, we start with the case $(N, m) = (1, 0)$, which also corresponds to a recently proposed model for antagonism in olfaction [41], with the role of k^{on} played by inverse affinity κ^{-1}, the role of τ played by efficiency η, and the spiking rate of olfactory receptor neurons a function $J(T_{N,m})$, that can be interpreted as a scoring function in the machine learning sense. In this case, $T_{1,0}$ simply computes the average quality τ_{avg} of ligands presented weighted by k_i^{on} (models with $N > m > 0$ give less intuitive results as will be shown in the following). It should be noted that while this computation is formally simple, biochemically it requires elaborated internal interactions, because a cell can not easily disentangle influence of individual receptors, see [11, 41] for explicit examples.

Starting from the computation of $\nabla_x J$ with respect to parameters k_i^{on} and τ_i, the FGSM perturbation is:

$$\eta = \epsilon \mathrm{sgn}\begin{pmatrix} \partial_{\tau_i} J \\ \partial_{k_i^{\mathrm{on}}} J \end{pmatrix} = \epsilon \mathrm{sgn}(A) \mathrm{sgn}\begin{pmatrix} k_i^{\mathrm{on}} \\ \tau_i - T_{1,0} \end{pmatrix}, \tag{1}$$

where $A = \frac{J'(T_{1,0})}{\sum k_i^{on}} > 0$. From the above expression, we find that an equivalent maximum adversarial perturbation is given by three simple rules (Fig. 2 A).

- Decrease all τ_i by ϵ
- Decrease k_i^{on} by ϵ for ligands with $\tau_i > T$
- Increase k_i^{on} by ϵ for ligands with $\tau_i < T$

The key relation to adversarial examples from [16] comes from considering what happens to the unbound receptors for which both k_i^{on} and τ_i are initially 0. Let us consider a situation with L identical bound ligands with $\{k_{\mathrm{on}} = 1, \tau\}$ giving response $T_{1,0}^{\mathrm{before}} = \tau$ where τ itself is of order 1 (in proper units). The three rules above imply that we are to decrease binding time by ϵ, and that all R previously unbound receptors are now to be bound by ligands with $k_{on} = \epsilon$ at small binding time ϵ. We compute the new response to be

$$T_{1,0}^{\mathrm{after}} = \frac{L(\tau - \epsilon) + \epsilon R \epsilon}{L + \epsilon R} = \frac{\tau - \epsilon + \frac{\epsilon R}{L}\epsilon}{1 + \frac{\epsilon R}{L}} \tag{2}$$

If there are many receptors compared to initial ligands, and assuming $\epsilon \ll \tau$, the relative change

$$\frac{T_{1,0}^{\mathrm{after}} - T_{1,0}^{\mathrm{before}}}{T_{1,0}^{\mathrm{before}}} \simeq -\frac{\frac{\epsilon R}{L}}{1 + \frac{\epsilon R}{L}} \tag{3}$$

is of order 1 when $\epsilon R \sim L$ giving a decrease comparable to the original response instead of being of order ϵ as we would naturally expect from small perturbations to all parameters. Thus, if a detection process is based on thresholding variable $T_{1,0}$, a significant decrease can happen with such perturbation, potentially shutting down response. Biologically, the limit where ϵR is big corresponds to a strong antagonistic effect of many weakly bound ligands, which yields the same effect as "competitive antagonism" in olfaction [41][2].

[1] an alternative choice without loss of generality is to consider a situation where for unoccupied receptors, k_i is 0 but τ_i is arbitrary, corresponding to a ligand available for binding

[2] One difference with olfaction is that for competitive antagonism, the concentration C is of order 1 while the

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

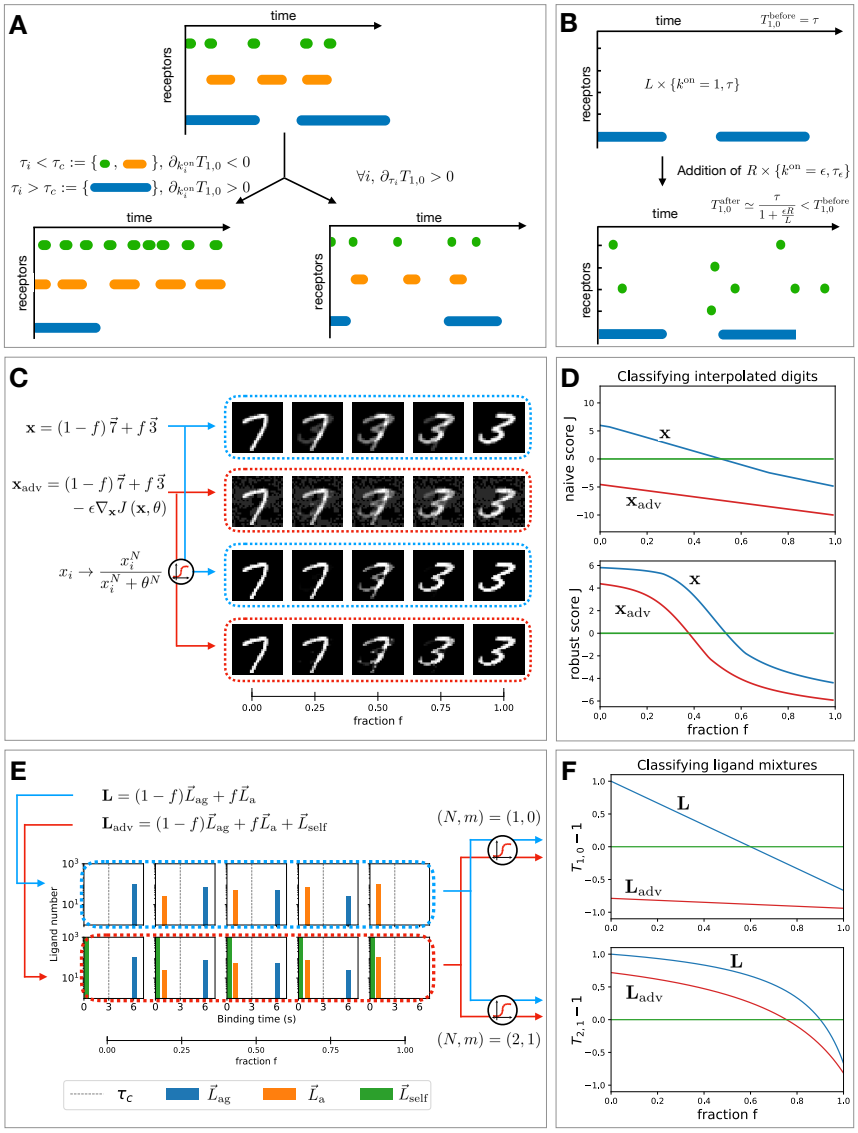


Figure 2. A) Schematics of FGSM applied to immune recognition. Decrease in weighted average $T_{1,0}$ comes from the given changes in ligand binding. B) Change of response upon addition of R ligands with small τ_ϵ. C) Interpolated digits with and without adversarial perturbation. Here, $\epsilon = 0.2$, $\theta = 0.5$, $N = 5$. D) Representation of the score for pictures of panel C for a naive neural network (top) and a robust neural network, which includes a biomimetic defence (bottom). The threshold is indicated by a green line at 0, top corresponds to 7 classification, bottom to 3 classification. E) Interpolated ligand mixtures with and without self ligands. Here, $(L_{ag}, \tau_{ag}) = (100, 6)$; $(L_a, \tau_a) = (100, 1)$; $(L_{self}, \tau_{self}) = (1000, 0.1)$ F) Representation of the score $T_{N,m} - 1$ for a naive immune classifier without proofreading with feedforward interaction (top) and a robust immune classifier which includes both proofreading and feedforward interaction (bottom). The threshold is indicated by a green line at 0, top corresponds to detection of agonists, bottom corresponds to no detection. In both cases, the naive networks interpolate the score linearly and are brittle to adversarial perturbation, while the score for robust networks is flatter, close to the initial sample and resistant to perturbation.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Behaviour across boundaries in sample space and adversarial perturbations

To further illustrate the correspondence, we compare the behaviour of a trained neural network classifying 3s and 7s with the adaptive proofreading model $(N, m) = (1, 0)$ for more general samples. We build linear interpolations between two samples on either side of the decision boundary for both cases (Fig. 2 C-F, dotted blue boxes, linear interpolation factor f varying between 0 and 1). This interpolation is the most direct way in sample spaces to connect objects in two different categories. The neural network classifies linearly interpolated digits, while the adaptive proofreading model classifies gradually changing ligand distributions.

We plot the output of the neural network x just before taking the final logistic function σ defined in Fig. 1 F and similarly, we plot $T_{N,m} - 1$ (in units rescaled by τ_c) for adaptive proofreading models. In both cases decision is thus based on the sign of the considered quantity. In the absence of adversarial/antagonistic perturbations, for both cases, we see that the score of the system almost linearly interpolates between values on either side of the classification boundary (top panel of Fig. 2 D, F, blue curves). However, in the presence of adversarial/antagonistic perturbations, the entire response is shifted way below the decision boundary (top panel of Fig. 2 D, F, red curves), so that in particular the initial samples at $f = 0$ (image of 7 or ligand distribution above threshold) are strongly misclassified.

Goodfellow et al. [16] proposed the linearity hypothesis as an explanation for this adversarial effect: adding $\eta = \epsilon \operatorname{sgn}(\nabla_x J)$ to the image leads to a significant perturbation on the scoring function J of order ϵd, with d the usually high dimensionality of the input space. Thus many weakly lit up "background" pixels in the initial image can conspire to fool the classifier, explaining the significant shift in the scoring function in Fig. 2 D top panel. This is consistent with the linearity we observe on the interpolation line even without adversarial perturbations. A more quantitative explanation based on averaging is given in [48] on a toy-model: after defining a label $y \in \{-1, +1\}$, a fixed probability p and a constant η, one can create a $(d+1)$–dimensional feature vector x

$$y \in \{-1, +1\}, \quad x_1 = \begin{cases} +y, & \text{w.p.} \quad p \\ -y, & \text{w.p.} \quad 1-p \end{cases}, \quad x_2, \ldots, x_{d+1} \in \mathcal{N}(\eta y, 1) \tag{4}$$

From this, we can built a 100% accurate classifier in the limit of $d \to \infty$ by averaging out the weakly correlated features $x_2, \ldots x_d$, which gives the score $f_{\text{avg}} = \mathcal{N}(\eta y, \frac{1}{d})$. Taking the sign of f_{avg} will coincide with the label y with 99% confidence for $\eta \geq 3/\sqrt{d}$. But such classification can be easily fooled by adding a small perturbation $\epsilon = -2\eta y$ to every component of the features, since it will shift the average by the same quantity $-2\eta y$, which can still be small if we take $\eta = \Theta(1/\sqrt{d})$.

We observe a very similar effect in the simplest adaptive proofreading model. The strong shift of the average $T_{1,0}$ in Eq. 2 is due to weakly bound receptors ϵR, which play the same role as the weak features (components $x_2, \ldots, x_{d+1}$ above), hiding the ground truth given by ligands of binding time τ (equivalent to x_1 above) to fool the classifier. We also see a similar linearity on the interpolation in Fig. 2 F top panel. There is thus a direct intuitive correspondence between adversarial examples in machine learning and the high number of available receptors R. In both cases, the change of scoring function (and corresponding misclassification) can be large despite the small amplitude ϵ of the perturbation. Once this perturbation is added, the system in Fig. 2 still interpolates between the two scores in a linear way, but with a strong shift due to the added perturbation.

Biomimetic defence for digit classification inspired by adaptive sorting

Kinetic proofreading [20, 38, 34] is a well-studied mechanism encoding different τ dependencies in the activation/repression branches of adaptive proofreading models [29]. The primary effect of kinetic proofreading is to non-linearly decrease the relative weight of weakly bound ligands with small binding times, thus ensuring defence against antagonism by weakly bound ligands. Inspired by this idea, we implement a simple defense for digit classification. Before feeding a picture to the neural network, we transform individual pixel values x_i of image $\mathbf{x}$ by

$$x_i \to \frac{x_i^N}{x_i^N + \theta^N} \tag{5}$$

affinity κ^{-1} is big, conversely, here the concentration R is big while k^{on} is low. Since we consider the product of both terms, both situations lead to similar effects, but our focus on a small change of k^{on} makes the comparison with machine learning more direct.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Similarly to the defence of adaptive proofreading where ligands with small τ are filtered out, this transformation squashes greyish pixels with values below threshold θ to black pixels, see Fig. 2 C bottom panels.

In Fig. 2 D, bottom panel, we show the improved robustness of the neural network armed with this defence. Here, the adversarial perturbation is filtered out efficiently. Strikingly, with or without adversarial perturbation, the score now behaves non-linearly along the interpolation line in sample space: it stays flatter over a broad range of f until suddenly crossing the boundary when the digit switches identity (even for a human observer) at $f = 0.5$. Similarly, for adaptive sorting with $(N, m) = (2, 1)$, antagonism is removed, and the score exhibits the same behaviour of flatness followed by a sudden decrease on the interpolation line. Thus, similar defence displays similar robust behaviour of the score in sample space.

Gradient dynamics identify two different regimes

The dynamics of the score along a trajectory in sample space can thus vary a lot as a function of the model considered. This motivates a more general study of a "worst case " scenario, i.e. gradient descent towards the decision boundary for different models. Krotov and Hopfield studied a similar problem for an MNIST digit classifier, encoded with generalized Rectified polynomials of variable degrees n [25] (reminiscent of the iterative FGSM introduced in [27]). The general idea is to find out how to most efficiently reach the decision boundary, and how this depends on the architecture of the decision algorithm. Krotov and Hopfield identified a qualitative change with increasing n, accompanied by a better resistance to adversarial perturbations [26, 25].

We consider the same problem for adaptive proofreading models, and study the dynamics of binding times for a ligand mixture when following the gradient of $T_{N,m}$ (akin to a potential in physics). The goal is to see how to most efficiently fool the decision-maker (or in biological terms, how to best antagonize it). We iteratively change the binding time of non agonist ligands $\tau < \tau_c$ to

$$\tau \to \tau - \epsilon \frac{\partial T_{N,m}}{\partial \tau} \tag{6}$$

while keeping the distribution of agonist ligands with $\tau > \tau_c$ constant. In the immune context, these dynamics can be thought of as a foreign agent trying to antagonize the immune system by rapidly mutating and generating antagonists ligands to mask its non-self part. Such antagonistic phenomena have been proposed as a mechanism for HIV escape [23, 35] and associated vaccine failure [21].

From a given ligand mixture with few ligands above threshold and many ligands below thresholds, we follow the dynamics of Eq. 6, and display the ligand distribution at the decision boundary for different values of N, m as well as the number of steps to reach the decision boundary in the descent defined by Eq. 6 (Fig. 3, see also Appendix 3 Figure 1 for another example with a visual interpretation). We observe two qualitatively different dynamics. For small m, we observe strong adversarial effects, as the boundary is almost immediately reached and the ligand distribution barely changes. As m increases, in Fig. 3 A the ligands in the distribution concentrate around one peak. For $m = 2$, a qualitative change occurs: the ligands suddenly spread over a broad range of binding times and the number of iterations in the gradient dynamics to reach the boundary drastically increases. For $m > 2$, the ligand distribution becomes bimodal, and the ligands close to $\tau = 0$ barely change, while a subpopulation of ligands peaks closer to the boundary. Consistent with this, the number of ϵ-sized steps to reach the boundary is 3 to 4 orders of magnitude higher for $m > 2$ as for $m < 2$.

Qualitative change in dynamics is due to a critical point

The qualitative change of behaviour observed at $m = 2$ can be understood by studying the contribution to the potential $T_{N,m}$ of ligands with very small binding times $\tau_\epsilon \sim 0$. Assuming without loss of generality that only two types of ligands are present (agonists $\tau_{\text{ag}} > \tau_{\text{c}}$ and spurious $\tau_{\text{spurious}} = \tau_\epsilon$), an expansion in τ_ϵ gives, up to a constant, $T_{N,m} \propto -\tau_\epsilon^m$ for small τ_ϵ (see Fig. 3 D for a representation of this potential and Appendix 3 for this calculation). In particular, for $0 < m < 1$, $\frac{\partial T_{N,m}}{\partial \tau_\epsilon} \propto -\tau_\epsilon^{m-1}$ diverges as $\tau_\epsilon \to 0$. This corresponds to a steep gradient of $T_{N,m}$ so that the system quickly reaches the boundary in this direction. The ligands close to $\tau_\epsilon \sim 0$ then quickly localize close to the minimum of this potential (unimodal distribution of ligand for small m on Fig. 3 A, B).

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

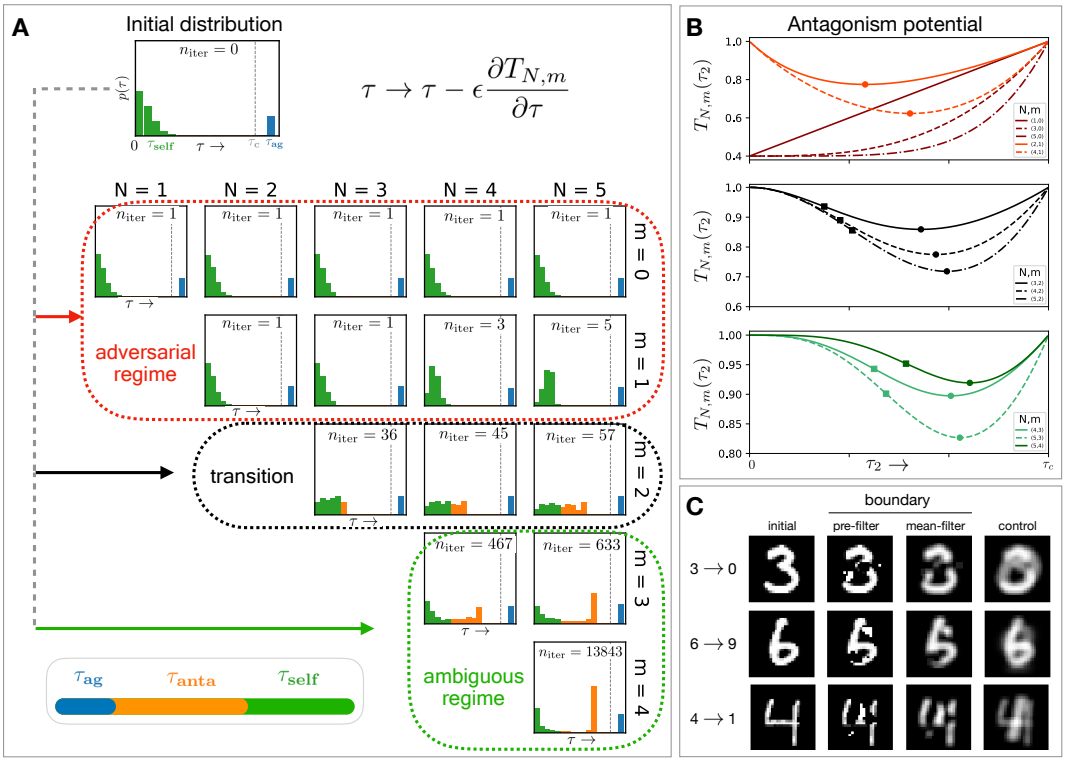


Figure 3. Characterization of the decision boundary following gradient descent dynamics. A) Ligand distribution at the decision boundary by applying iterative gradient descent (top right of the panel) to an initial distribution (top left). For various cases (N, m) we change the binding time of self ligands along the steepest gradient until reaching the decision boundary. n_{iter} indicates the number of iterations needed to reach the decision boundary. B) $T_{N,m}$ for mixtures of ligands at τ_c with ligands at τ_2, as a function of τ_2 for various (N, m). Antagonism strength is maximal when $T_{N,m}$ is minimal. Minima and inflexion points are indicated with a circle and square. C) Few-pixel attack as a way of circumventing proofreading or local contrast defence, while creating ambiguous digits. We add a 3x3 mean-filter to demonstrate the ambiguity of digits at the decision boundary. The control image is the mean filtered initial digit combined with the locally contrasted average target digit. Note that the control is also lacking a clear ground truth.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The potential close to $\tau_\epsilon \sim 0$ flattens for $1 < m < 2$, but it is only at $m = 2$ that a critical point appears at $\tau_\epsilon = 0$, qualitatively modifying the dynamics defined by Eq. 6. For $m \geq 2$, due to the new local flatness of this gradient, ligands at $\tau = 0$, the critical point of Eq. 6, are pinned by the dynamics. By continuity, dynamics of the ligands slightly above $\tau_\epsilon = 0$ are critically slowed down, making it much more difficult for them to reach the boundary. This explains both the sudden broadening of the ligand distribution, and the associated increase in the number of steps to reach the decision boundary. Conversely, an inflexion point (square) appears in between the minimum (circle) and $\tau_\epsilon = 0$ (Fig 3 B). Ligands close to the inflexion point separate and move more quickly towards the minimum of potential, explaining the bimodality at the boundary. For larger (N, m) we obtain flatter potentials, and a larger number of iterations. In Appendix 2, we further describe the consequence of adding proofreading steps on the position of the boundary itself, using another concept of machine learning called "boundary tilting" [46] (Appendix 2, Figure 1 and Table 1).

Categorization of attacks

The transition at $m = 2$ is strongly reminiscent of the transition observed by Krotov and Hopfield in their study of gradient dynamics similar to Eq. 6 [25]. In both our works, we see that there are (at least) two kinds of attacks that can bring samples to the decision boundary. The FGSM corresponds to small perturbations to the input in terms of L_∞ norm leading to modifications of many background pixels in [25] or many weakly bound ligands for the adaptive proofreading case, also similar to the meaningless changes in meta-features described above in Eq. 4 [48].

Defence against the FGSM perturbation is implemented through a higher degree n of the rectified polynomials in [25], while in adaptive proofreading, this is done through critical slowing down of the dynamics of Eq. 6 for $m > 2$. The latter models are nevertheless sensitive to another kind of attack with many fewer perturbations of the inputs but with bigger magnitude. This corresponds to digits at the boundary where few well-chosen pixels are turned on in [25]. For adaptive proofreading models this leads to the ligand distribution becoming bimodal at the decision boundary. Two important features are noteworthy. First, the latter perturbations are difficult to find through gradient descent (as illustrated by the many steps to reach the boundary in Fig. 3 A). Second, the perturbations appear to be "meaningful" and it is difficult or even impossible to recover the "ground truth" by inspecting the sample at the decision boundary. Digits at the boundary for [25] appear indeed ambiguous to a human observer, and ligand distribution peaking just below threshold are potentially misinterpreted biologically due to inherent noise. This has actually been observed experimentally in T cells, where strong antagonists are also weak agonists [2, 13], meaning that T cells do not take reliable decisions in this regime.

Biomimetic defenses against few-pixel attacks

It is then worth testing the sensitivity to localized stronger attacks of digit classifiers, helped again with biomimetic defences. The natural analogy is to implement attacks based on strong modification of few pixels [43]. For this problem, we choose to implement a two-tier biomimetic defence: we implement first the transformation defined in Eq. 5, that will remove influence of the FGSM types of perturbations by flattening the local landscape as in Fig. 2 D. In addition, we choose to add a second layer of defence where we simply average out locally pixel values. This can be interpreted biologically as a process of receptor clustering or time-averaging. Time-averaging has been shown to be necessary in a stochastic version of adaptive proofreading [13, 29], where temporal intrinsic noise would otherwise make the system cross the boundary back and forth endlessly. In the machine learning context, local averaging has been recently proposed as a way to defend against few pixel attacks [52], which thus can be considered as the analogous of biochemical noise.

We then train multiple classifiers between different pairs of handwritten digits. Following the approach of "one pixel" attack [43], we consider digits classified in presence of this two-tier defence, then sequentially fully turn pixels on or off ranked by their impact on the scoring function, until we reach the decision boundary.

Representative results of such few-pixels attacks with biomimetic defences are illustrated in Fig 3 C ("pre-filter" column, local average is shown for further comparison in column "mean-filter"), with other examples shown in Appendix 4 Figure 1 and details on the behaviour of scoring functions in

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix 4 Figure 2. Clearly the attacked samples at the boundaries hide the "ground truth" of the initial digit, and as such can not be considered are meaningless adversarial perturbations. Samples at the boundary superficially look like printed Japanese "kanas" or Greek characters (e.g. attacks from 0 to 1 typically look like a ϕ, see Appendix 4), making them impossible to classify as Arabic digits even for a human observer. This is consistent with the ambiguous digits observed for big n by Krotov and Hopfield [25]. In other cases, samples at the boundary between two digits actually look like another digit: for instance, we see that the sample at the boundary between a 6 and an 9 look like a 5 (or a Japanese ち). This observation is consistent with previous work attempting to interpolate in latent space between digits [4], where at the boundary a third digit corresponding to another category may appear. We also compare in Fig 3 C the sample seen by the classifier at the boundary after the biomimetic defences with a "control" corresponding to the average between the initial digit and the target of the attack (corresponding to the interpolation factor $f = 0.5$ in Fig. 2 C–D). It is then quite clear that the sample generated by the attack is rather close to this control boundary image. This, combined with the fact samples at the boundary still look like printed characters without clear ground truth indicate that the few pixel attacks implemented here actually select for "meaningful" features. The existence of meaningful features in the direction of the gradient have been identified as a characteristic of networks robust to adversarial perturbation [48] similar to results of [25] and our observation for adaptive proofreading models above.

Discussion

Complex systems (*in vivo* or *in silico*) integrate sophisticated decision making processes. Our work illustrates common features between neural networks and a general class of adaptive proofreading models, especially with regards to mechanisms of defence against targeted attacks. Parallels can be drawn between these past approaches, since the models of adaptive proofreading presented here were first generated with *in silico* evolution aiming at designing immune classifiers [29]. Strong antagonism naturally appeared in the simplest simulations, and required modification of objective functions very similar to adversarial training [16].

Through our analogy with adaptive proofreading, we are able to identify the presence of a critical point (due to kinetic proofreading) as the crucial mediator of robust adversarial defense, essentially squashing the spurious adversarial directions. Another layer of defence can be added with local averaging. This is in line with current research on adversarial robustness, showing that robust networks exhibit a flat loss landscape near each training sample [36].

An interesting by-product of local flatness is the appearance of an inflexion point in the gradient descent dynamics. If the scoring function is flat close to a sample far from the boundary, nonzero at the boundary, then flat close to another sample, an inflexion point is expected via Rolle's theorem. This is visible in Fig. 2 D–F: while the score of non-robust classifiers is linear when moving towards the decision boundary, the scoring function of classifiers resistant to adversarial perturbations is flat at $f = 0$ and only significantly changes when the input becomes ambiguous near the inflexion point. The presence of this inflexion point is bound to strongly influence the gradient descent dynamics. For instance, for adaptive proofreading models, the ligand distribution following the dynamics of Eq. 6 changes from unimodal to bimodal at the boundary, creating ambiguous samples. For a robust classifier, such samples are thus expected to appear close to the decision boundary since they coincide with large gradients. As such they could correspond to meaningful features (contrasting the meaningless adversarial perturbations), as we show in Fig. 3 with our digit classifier with biomimetic defence. Examples in image classification might include the perturbed animal pictures fooling humans [8] with chimeric images that combine different animal parts (such as spider and snake) or the meaningful adversarial transformations between samples found in [48]. Similar properties have been observed experimentally for ambiguous samples in immune recognition: maximally antagonizing ligands have a binding time just below the decision threshold [2]. We interpret this property as a consequence of the flat landscape far from the decision threshold leading to a steep gradient close to it [13, 12].

A possible caveat of our analogy is that a clear decision axis in the τ direction can be accounted for explicitly in adaptive proofreading models (even though it is still convolved with ligand quantity, thus requiring an adaptive proofreading architecture). In machine learning the space of inputs is much more complex, and there are generally more than two categories. Here, the algorithm effectively has to learn representations, such as pixel statistics and spatial correlations in images [24]. However, underlying, low manifold descriptions could locally still combine higher level information in ways similar to parameter τ, so that the theory presented here could still apply once those directions are discovered. Coming back to biology, it was shown mathematically that for the classification problem

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of discriminating ligand quality irrespective of their quantity, one always gets antagonism close to the boundary [12]. This is precisely due to the necessary presence of a significant gradient in the direction of the decision-making. However, one can change the binding time of maximally antagonizing ligands via the nonlinearities in kinetic proofreading. Similar inevitability theorems might be generalizable to machine learning, for instance via a robustness-accuracy trade-off [48]. Adversarial examples are potentially impossible to fully remove, yet the effective adversarial perturbation may shift from a pile of meaningless features to a combination of meaningful features, giving ambiguous patterns at the decision boundary.

From the biology standpoint, new insights may come from the general study of computational systems built via machine learning. Our study of Fig. 3, inspired by gradient descent in machine learning, suggests that cellular decision-makers exist in two regimes. The difference between these regimes are geometric by nature, with or without critical points. The case $m < 2$ with a steep gradient could be more relevant in signalling contexts to separate mixtures of inputs, so that every weak perturbation *should* be detected. For olfaction it has been suggested that strong antagonism allows for a "rescaling" of the distribution of typical odor molecules, ensuring a broad range of detection irrespective of the quantity of molecules presented [41]. The case $m \geq 2$ is much more resistant to adversarial perturbations, and could be most relevant in an immune context where T cells filter out antagonistic perturbations. This might be relevant for the pathology of HIV infections [23, 35, 21] or, more generally, could provide explanations on the diversity of altered peptide ligands [49]. We also expect similar classification problems to occur at the population-level, e.g. when T cells interact with each other to refine individual immune decision-making [5, 50].

We have connected machine learning algorithms to models of cellular decision-making, and in particular their defence strategies against "adversarial" attacks. More defences against adversarial examples might be found in the real world, for instance in biofilm-forming in bacteria [53], in size estimation of animals [28], or might be needed for proper detection of physical 3D objects [3] and road signs [9]. Understanding the whole range of possible antagonistic perturbations may also prove crucial for describing immune defects, including immune escape of cancer cells. It is thus important to further clarify possible scenarios for fooling the classifier in both artificial and living systems.

Acknowledgements

We thank Joelle Pineau and members of the François group for useful discussions. P.F. is supported by a Simons Investigator in Mathematical Modelling of Living Systems award, an Integrated Quantitative Biology Initiative award and a Natural Sciences and Engineering Research Council award (Discovery Grant). T.J.R. receives funding from the Centre for Applied Mathematics in Bioscience and Medicine (graduate award) and McGill Physics (Schulich award). E.B. acknowledges support from the Samsung Advanced Institute of Technology and the Fonds de Recherche du Québec – Nature et Technologies (graduate award).

References

[1] **Akhtar N**, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. arXiv preprint arXiv:180100553. 2018; .

[2] **Altan-Bonnet G**, Germain RN. Modeling T cell antigen discrimination based on feedback control of digital ERK responses. PLOS Biology. 2005; 3(11):e356.

[3] **Athalye A**, Engstrom L, Ilyas A, Kwok K. Synthesizing Robust Adversarial Examples. In: *Proceedings of the 35th International Conference on Machine Learning*, vol. 80; 2018. p. 284–293.

[4] **Berthelot D**, Raffel C, Roy A, Goodfellow I. Understanding and Improving Interpolation in Autoencoders via an Adversarial Regularizer. arXiv preprint arXiv:180707543. 2018; .

[5] **Butler TC**, Kardar M, Chakraborty AK. Quorum sensing allows T cells to discriminate between self and nonself. Proceedings of the National Academy of Sciences. 2013; 110(29):11833–11838.

[6] **Chakraborty AK**, Weiss A. Insights into the initiation of TCR signaling. Nature immunology. 2014; 15(9):798.

[7] **Dittel BN**, Germain RN, Janeway Jr CA, et al. Cross-antagonism of a T cell clone expressing two distinct T cell receptors. Immunity. 1999; 11(3):289–298.

[8] **Elsayed G**, Shankar S, Cheung B, Papernot N, Kurakin A, Goodfellow I, Sohl-Dickstein J. Adversarial Examples that Fool both Computer Vision and Time-Limited Humans. In: *Advances in Neural Information Processing Systems*; 2018. p. 3911–3921.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[9] **Eykholt K**, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, Prakash A, Kohno T, Song D. Robust Physical-World Attacks on Deep Learning Visual Classification. In: *The IEEE Conference on Computer Vision and Pattern Recognition*; 2018. p. 1625–1634.

[10] **Feinerman O**, Germain RN, Altan-Bonnet G. Quantitative challenges in understanding ligand discrimination by α β T cells. Molecular Immunology. 2008; 45(3):619.

[11] **François P**, Altan-Bonnet G. The case for absolute ligand discrimination: modeling information processing and decision by immune T cells. Journal of Statistical Physics. 2016; 162(5):1130–1152.

[12] **François P**, Hemery M, Johnson KA, Saunders LN. Phenotypic spandrel: absolute discrimination and ligand antagonism. Physical Biology. 2016; 13(6):066011.

[13] **François P**, Voisinne G, Siggia ED, Altan-Bonnet G, Vergassola M. Phenotypic model for early T-cell activation displaying sensitivity, specificity, and antagonism. Proceedings of the National Academy of Sciences. 2013; p. 201300752.

[14] **Gascoigne NR**, Zal T, Alam SM. T-cell receptor binding kinetics in T-cell development and activation. Expert Reviews in Molecular Medicine. 2001; 3(6):1–17.

[15] **Germain RN**, Stefanová I. The dynamics of T cell receptor signaling: complex orchestration and the key roles of tempo and cooperation. Annual review of immunology. 1999; 17(1):467–522.

[16] **Goodfellow IJ**, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. arXiv preprint arXiv:14126572. 2014; .

[17] **Govern CC**, Paczosa MK, Chakraborty AK, Huseby ES. Fast on-rates allow short dwell time ligands to activate T cells. Proceedings of the National Academy of Sciences. 2010; p. 201000966.

[18] **Grosse K**, Manoharan P, Papernot N, Backes M, McDaniel P. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:170206280. 2017; .

[19] **Hinton G**, Deng L, Yu D, Dahl GE, Mohamed Ar, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine. 2012; 29(6):82–97.

[20] **Hopfield JJ**. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proceedings of the National Academy of Sciences. 1974; 71(10):4135–4139.

[21] **Kent SJ**, Greenberg PD, Hoffman MC, Akridge RE, McElrath MJ. Antagonism of vaccine-induced HIV-1-specific CD4+ T cells by primary HIV-1 infection: potential mechanism of vaccine failure. The Journal of Immunology. 1997; 158(2):807–815.

[22] **Kersh GJ**, Kersh EN, Fremont DH, Allen PM. High-and low-potency ligands with similar affinities for the TCR: the importance of kinetics in TCR signaling. Immunity. 1998; 9(6):817–826.

[23] **Klenerman P**, Rowland-Jones S, McAdam S, Edwards J, Daenke S, Lalloo D, Köppe B, Rosenberg W, Boyd D, Edwards A, et al. Cytotoxic T-cell activity antagonized by naturally occurring HIV-1 Gag variants. Nature. 1994; 369(6479):403.

[24] **Krizhevsky A**, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*; 2012. p. 1097–1105.

[25] **Krotov D**, Hopfield J. Dense associative memory is robust to adversarial inputs. Neural computation. 2018; p. 1–17.

[26] **Krotov D**, Hopfield JJ. Dense associative memory for pattern recognition. In: *Advances in Neural Information Processing Systems*; 2016. p. 1172–1180.

[27] **Kurakin A**, Goodfellow I, Bengio S. Adversarial machine learning at scale. arXiv preprint arXiv:161101236. 2016; .

[28] **Laan A**, de Polavieja G. Sensory cheating: adversarial body patterns can fool a convolutional visual system during signaling. bioRxiv. 2018; p. 326652.

[29] **Lalanne JB**, François P. Principles of adaptive sorting revealed by in silico evolution. Physical Review Letters. 2013; 110(21):1–5.

[30] **Lalanne JB**, François P. Chemodetection in fluctuating environments: Receptor coupling, buffering, and antagonism. Proceedings of the National Academy of Sciences. 2015; 112(6):1898–1903.

[31] **LeCun Y**, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436.

[32] **LeCun Y**, Cortes C. The MNIST database of handwritten digits. . 1998; .

[33] **Lever M**, Lim HS, Kruger P, Nguyen J, Trendel N, Abu-Shah E, Maini PK, van der Merwe PA, Dushek O. Architecture of a minimal signaling pathway explains the T-cell response to a 1 million-fold variation in antigen affinity and dose. Proceedings of the National Academy of Sciences. 2016; 113(43):E6630–E6638.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[34] **McKeithan TW**. Kinetic proofreading in T-cell receptor signal transduction. Proceedings of the National Academy of Sciences. 1995; 92(11):5042–5046.

[35] **Meier UC**, Klenerman P, Griffin P, James W, Köppe B, Larder B, McMichael A, Phillips R. Cytotoxic T lymphocyte lysis inhibited by viable HIV mutants. Science. 1995; 270(5240):1360–1362.

[36] **Moosavi-Dezfooli SM**, Fawzi A, Uesato J, Frossard P. Robustness via curvature regularization, and vice versa. arXiv preprint arXiv:181109716. 2018; .

[37] **Mora T**. Physical limit to concentration sensing amid spurious ligands. Physical review letters. 2015; 115(3):038102.

[38] **Ninio J**. Kinetic amplification of enzyme discrimination. Biochimie. 1975; 57(5):587–595.

[39] **Papernot N**, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*; 2017. p. 506–519.

[40] **Pedregosa F**, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

[41] **Reddy G**, Zak JD, Vergassola M, Murthy VN. Antagonism in olfactory receptor neurons and its implications for the perception of odor mixtures. eLife. 2018; 7:e34958.

[42] **Siggia ED**, Vergassola M. Decisions on the fly in cellular sensory systems. Proceedings of the National Academy of Sciences. 2013; 110(39):E3704–E3712.

[43] **Su J**, Vargas DV, Kouichi S. One pixel attack for fooling deep neural networks. arXiv preprint arXiv:171008864. 2017; .

[44] **Sutskever I**, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: *Advances in Neural Information Processing Systems*; 2014. p. 3104–3112.

[45] **Szegedy C**, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv preprint arXiv:13126199. 2013; .

[46] **Tanay T**, Griffin L. A boundary tilting persepective on the phenomenon of adversarial examples. arXiv preprint arXiv:160807690. 2016; .

[47] **Torigoe C**, Inman JK, Metzger H. An unusual mechanism for ligand antagonism. Science. 1998; 281(5376):568–572.

[48] **Tsipras D**, Santurkar S, Engstrom L, Turner A, Madry A. Robustness may be at odds with accuracy. arXiv preprint arXiv:180512152. 2018; 1.

[49] **Unanue ER**. Altered peptide ligands make their entry. The Journal of Immunology. 2011; 186(1):7–8.

[50] **Voisinne G**, Nixon BG, Melbinger A, Gasteiger G, Vergassola M, Altan-Bonnet G. T cells integrate local and global cues to discriminate between structurally similar antigens. Cell reports. 2015; 11(8):1208–1219.

[51] **Wong E**, Kolter Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In: *International Conference on Machine Learning*; 2018. p. 5283–5292.

[52] **Xie C**, Wu Y, van der Maaten L, Yuille A, He K. Feature Denoising for Improving Adversarial Robustness. arXiv preprint arXiv:181203411. 2018; .

[53] **Yan J**, Deforet M, Boyle KE, Rahman R, Liang R, Okegbe C, Dietrich LE, Qiu W, Xavier JB. Bow-tie signaling in c-di-GMP: Machine learning in a simple biochemical network. PLOS Computational Biology. 2017; 13(8):e1005677.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix 1

Mathematical description of the adaptive proofreading models

We assume an idealized situation where a given receptor i, upon ligand binding (on-rate k_i^{on}) can exist in N biochemical states (corresponding to phosphorylation stages of the receptor tails in the immune context [34, 22]). Those states allow the receptor to effectively compute different quantities, such as $c_m^i = k_i^{\mathrm{on}}\tau_i^m, 0 \leq m \leq N$, which can be done with kinetic proofreading [20, 38, 34]. In particular, ligands with larger τ give a relatively larger value of c_N^i due to the geometric amplification associated with proofreading steps. We assume receptors to be identical, so that any downstream receptor processing by the cell must be done on the sum $C_m = \sum_i c_m^i = \sum_i k_i^{on}\tau_i^m$. We also consider a quenched situation in which only one ligand is locally available for binding to every receptor. In reality, there is a constant motion of ligands, such that k_i^{on} and τ_i are functions of time and stochastic treatments are required [42, 30, 37], but on the time-scale of primary decision-making it is reasonable to assume that the ligand distribution does not change much [2]. Probability of decision-making in this context is a monotonically increasing function of the quantity

$$T_{N,m} = \frac{\sum_i k_i^{\mathrm{on}}\tau_i^N}{\sum_i k_i^{\mathrm{on}}\tau_i^m}. \tag{7}$$

If L ligands with identical τ and k^{on} are presented to the T cell, we have $T_{N,m} = \frac{k^{\mathrm{on}}L\tau^N}{k^{\mathrm{on}}L\tau^m} = \tau^{N-m}$. In this situation,a threshold ($T_{N,m} > \tau_c^{N-m}$) can be easily defined for decision, irrespective of the quantity L of ligands presented.

If we now add L_{a} antagonists with lower binding time $\tau_{\mathrm{a}} < \tau$ and equal on-rate k^{on}, we have $T_{N,m} = \frac{L\tau^N + L_{\mathrm{a}}\tau_{\mathrm{a}}^N}{L\tau^m + L_{\mathrm{a}}\tau_{\mathrm{a}}^m}$, which is smaller than the response τ^{N-m} for a single type of ligands, corresponding to ligand antagonism (Fig. 1 D, main text) [15, 7, 2, 11]

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

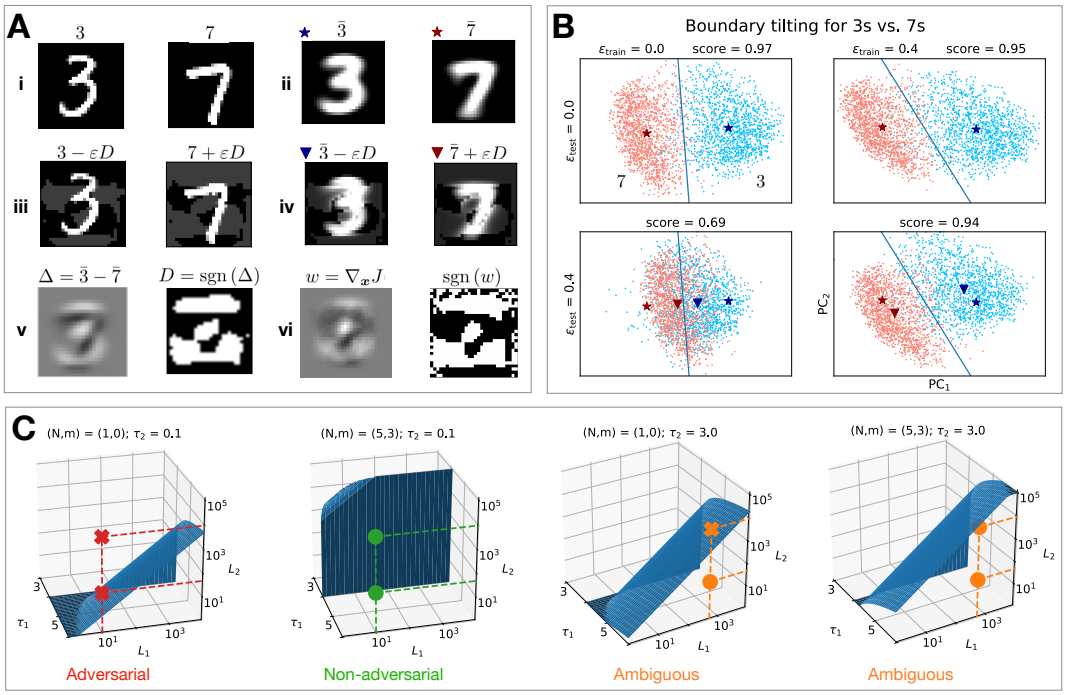
Appendix 2

Boundary tilting

To further draw the connection between machine learning and adaptive proofreading models, we will study a framework to interpret adversarial examples called boundary tilting [46]. We will first illustrate this effect on the discrimination of the original MNIST 3 vs 7 problem MNIST from [16]), after which we will interpret boundary tilting via proofreading.

Digit classification

A typical 3 and 7 (i), the averages $\bar{3}$ and $\bar{7}$ (ii), and the corresponding adversarial examples (iii, iv) are shown in Fig. 1 A. Tanay and Griffin [46] pointed out that the adversarial perturbation generated with the Fast Gradient Sign Method (FGSM) proposed in [16] can also be found via $D = \text{sign}\,(\bar{3} - \bar{7})$, Fig. 1 A (v). Note the similarity to the adversarial perturbation from the FGSM $\text{sign}(w) = \text{sign}\,(\nabla_x J)$ (Fig. 1 A (vi)).



Appendix 2 Figure 1. Boundary tilting in one-dimensional digit classification. A) (i) Typical 3 and 7 from MNIST. (ii) Average 3, 7 of the traditional test set, (iii,iv) with adversarial perturbation, found by (v) subtracting the sign of $\bar{3}$ from $\bar{7}$, which corresponds to (vi), the perturbation found with FGSM B) Projection of the digits on the first principal components. The classes are separated by a linear Support Vector Classifier (blue), the average of the classes with and without adversarial perturbation is shown by the triangle and star. We have cycled through permutations of adversarial training and/or adversarial testing. Note how the boundary tilts on the right panels, and how the triangle moves parallel to the decision boundary. C) Decision boundary of the immune model. The region under the surface is the response regime, the region above is the no-response regime. The classifier with a single proofreading step $(N, m) = (1, 0)$ fails to observe agonists in three of the four marked mixtures, while the robust classifier $(N, m) = (5, 3)$ correctly responds to each indicated mixture.

To reveal the linearity of binary digit discrimination, we computed the principal components (PCs) of the traditional training set of 3s and 7s, and projected all digits in the test set on PC_1 and PC_2 (Fig. 1 B). With a linear Support Vector Classifier (ordinary linear regression) trained on the transformed coordinates PC_1 and PC_2 of the training set, we achieve over 95% accuracy in the test set. While such accuracy is far from the state-of-the-art in digit recognition, it is much higher than typical detection accuracy for single cells (e.g. T cells present false negative rates of 10 % for strong antagonists [2]). The red and blue star in Fig. 1 denote the average digit $\bar{3}, \bar{7}$.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Next, we transformed the test set as $3 \to 3' = 3 - \epsilon_{\text{test}} D, \quad 7 \to 7' = 7 + \epsilon_{\text{test}} D$, where $\epsilon_{\text{test}} = 0.4$ is the strength of the adversarial perturbation (Fig. 1 A (iii)). $\bar{3}'$ and $\bar{7}'$ moved closer in Fig. 1 B, orthogonal to the decision boundary and along the line between the initial averages. This adversarial perturbation moves the digits in what we call an adversarial direction perpendicular to the decision boundary, and reduces the accuracy of the linear regression model to a mere 69%.

Goodfellow et al. proposed adversarial training as a method to mitigate adversarial effects by FGSM. We implemented adversarial training by adding the adversarial perturbation $\epsilon_{\text{train}} D_{\text{train}} = \epsilon_{\text{train}}(\bar{3}_{\text{train}} - \bar{7}_{\text{train}})$ to the images in the training set, computing the new PCs and training the linear regression model. This effectively "tilts" the decision boundary, while preserving 95% accuracy. In the presence of the original adversarial perturbations, we see the effect of the tilted boundary: the perturbation moves digits parallel along the decision boundary, which results in good robust accuracy. This is a illustration example of the more general phenomenon studied in [46].

Boundary tilting and categorizing perturbations

We consider the change in $T_{N,m}$ for arbitrary N, m upon addition of many spurious ligands. Generalizing Eq. 2 in the main text gives

$$T_{N,m}^{\text{after}} = \frac{L(\tau - \epsilon)^N + \epsilon R \tau_\epsilon^N}{L\tau^m + \epsilon^m R} = \frac{(\tau - \epsilon)^N + \epsilon \frac{\epsilon^N R}{L} \tau_\epsilon}{\tau^m + \frac{\epsilon^{m+1} R}{L}}. \tag{8}$$

From this expression, we note that $T_{N,m}$ is changing significantly with respect to its initial value upon addition of many weakly bound ligands as soon as $\epsilon^{m+1} R$ is of order L. Thus, the effect described in the main text for weighted averages where $(N, m) = (1, 0)$ also holds for nonlinear computations as long as m is small. It appears that the general strategy to defend against this adversarial perturbation is by increasing m, as previously observed in [29]. Biochemically, this is done with kinetic proofreading [34, 2, 13], i.e. we take an output $T_{N,m}$ with $N > m \geq 1$. Here, the output is no longer sensitive to the addition of many weakly bound self ligands, yielding an inversion of the antagonistic hierarchy where the strongest antagonizing ligands exist closer to threshold [12]. An extreme case has been proposed for immune recognition where the strongest antagonists are found just below the threshold of activation [2].

We numerically compute how the decision boundary changes when L_2 ligands at τ_2 are added to the initial L_1 ligands at τ_1, i.e. we compute the manifold so that

$$T_{N,m}(\{L_1, \tau_1; L_2, \tau_2\}) = \frac{\tau_1^N L_1 + \tau_2^N L_2}{\tau_1^m L_1 + \tau_2^m L_2} \tag{9}$$

is equal to $T_{N,m}(\{L_1, \tau_c\}) = \tau_c^{N-m}$. We represent this boundary for fixed τ_2 and variable L_1, L_2, τ_1 in Fig. 1 C. Boundary tilting is studied with respect to the reference $L_2 = 0$ plane corresponding to the situation of pure L_1 ligands at τ_1, where the boundary is the line $\tau_1 = \tau_c$. The case $(N, m) = (1, 0)$ (Fig. 1 C, left panel), corresponds to a very tilted boundary, close to the plane $L_2 = 0$, and a strong antagonistic case. In this situation, assuming $\tau_1 \sim \tau_c$, each new ligand added with τ_2 close to 0 gives a reduction of $T_{1,0}$ proportional to $\frac{\tau_c}{L_1}$ in the limit of small L_2 (see next section, [11]), which is again of the order of the response $T_{1,0} = \tau_1 \sim \tau_c$ in the plane $L_2 = 0$. This is clearly not infinitesimal, corresponding to a steep gradient of $T_{1,0}$ in the L_2 direction. We call the perturbation in this case "adversarial". This should be contrasted to the case for higher m (Fig. 1 C, middle left) where the boundary is vertical, independent of L_2, such that decision-making is based only on the initially present L_1 ligands at τ_1. Here, the change of response induced by the addition of each ligand with small binding time τ_2 is τ_2^m, due to proofreading a very small number when $\tau_2 \simeq 0$ [11]. Contrary to the previous case, the gradient of $T_{N,m}$ with respect to this vertical direction is almost flat and very small compared to the response in the $L_2 = 0$ plane. We call the perturbation in this case "non-adversarial".

Tilting of the boundary only occurs when τ_2 gets sufficiently close to the threshold binding time τ_c (Fig. 1 C, right panels). In this regime, each new ligand added with quality $\tau_2 = \tau_c - \epsilon$ contributes an infinitesimal change of $T_{N,m}$ proportional to $\frac{\tau_c - \tau_2}{L_1} = \epsilon / L_1$, which gives a weak gradient in the direction L_2. But even with such small perturbations one can easily cross the boundary because of the proximity of τ_2 to τ_c, which explains the tilting. The cases where the boundary is tilted and the gradient is weak are of a different nature compared to the adversarial case of Fig. 1 C, left panel. Here the boundary is tilted as well, but the gradient is steep, not weak. For this reason we term the cases on the right panels

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

"ambiguous". Similar ambiguity is observed experimentally: it is well known that antagonists (ligands close to thresholds) also weakly agonize an immune response [2]. Our categorization of perturbations is presented in Table 1. Scripts for boundary tilting in ligand discrimination and digit discrimination are available at https://github.com/tjrademaker/advxs-antagonism-figs/.

Appendix 2 Table 1. Categories of perturbations

	Boundary tilting	Gradient when adding one antagonistic ligand
Adversarial	yes	steep ($\mathcal{O}(1)$)
Non-adversarial	no	almost flat ($\mathcal{O}(\epsilon^m)$)
Ambiguous	yes	weak ($\mathcal{O}(\epsilon)$)

Gradient in the L_2 direction

We recall results from [12] to show how the addition of subthreshold ligands one at a time changes the output. We first consider $\{L_1, \tau_c\}$ threshold ligands with output

$$T_{N,m}(L_1, \tau_c) = \tau_c^{N-m}. \tag{10}$$

The main result of [12] is the linear response of $T_{N,m}(L_1, \tau_c)$ to the addition of $\{L_2, \tau_c - \epsilon\}$ subthreshold ligands.

$$T_{N,m}\left(\{L_1, \tau_c; L_2, \tau_c - \epsilon\}\right) = T\left(L_1 + L_2, \tau_c\right) - \epsilon L_2 \mathcal{A}\left(L_1 + L_2, \tau_c\right) \tag{11}$$

$$= \tau_c^{N-m} - \epsilon \frac{L_2}{L_1 + L_2} \frac{d}{d\tau} T_{N,m}(L_1 + L_2, \tau)\Big|_{\tau=\tau_c}, \tag{12}$$

where we used the definition

$$\mathcal{A}\left(L, \tau_c\right) = \frac{1}{L} \frac{d}{d\tau} T_{N,m}(L, \tau)\Big|_{\tau=\tau_c}. \tag{13}$$

for the coefficient in a mean-field description. As the derivative $\frac{d}{d\tau} T_{N,m}(L, \tau)\Big|_{\tau=\tau_c} > 0$, and $\epsilon = \tau_2 - \tau_c$, each additional subthreshold ligand at τ_2 decreases the output with a value proportional to

$$\frac{\tau_c - \tau_2}{L_1}. \tag{14}$$

In the case $(N, m) = (1, 0)$, the mean-field approximation is exact, i.e. the first derivative of $\frac{dT}{d\tau}$ is the only nonzero derivative, given by

$$\mathcal{A}(L_1, \tau_c) = \frac{1}{L_1} \frac{d}{d\tau} \tau\Big|_{\tau=\tau_c} = \frac{1}{L_1}. \tag{15}$$

With the addition of a single subthreshold ligand $\tau_2 \sim 0$, so that $\epsilon \sim \tau_c$, the output is maximally reduced by $\frac{\tau_c}{L_1+1} \simeq \frac{\tau_c}{L_1}$, a finite quantity, as described in the main text. For higher m, the linear approximation holds only for ligands at τ_2 close to threshold.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix 3

Ligand distribution at the decision boundary

Adaptive proofreading is well-suited to characterize the decision boundary between two classes, because we can work with an analytical description. We want to know how to most efficiently change the binding time of the spurious binding ligand (with small τ) to cause the model to reach the decision boundary. We have taken inspiration from [25] and adapted our approach from the iterative FGSM [27]. At first, we sample L_{s} self ligands from a normal distribution folded around $\tau = 0$ and L_{Ag} agonist ligands from a narrowly peaked normal distribution above τ_c. The agonist ligand distribution, the "signal" in the immune picture, remains constant. Next, we bin ligands in M equally spaced bins τ_b, $b \in [1, M]$, and we compute the gradient for those bins for which $\tau_b < \tau_c$

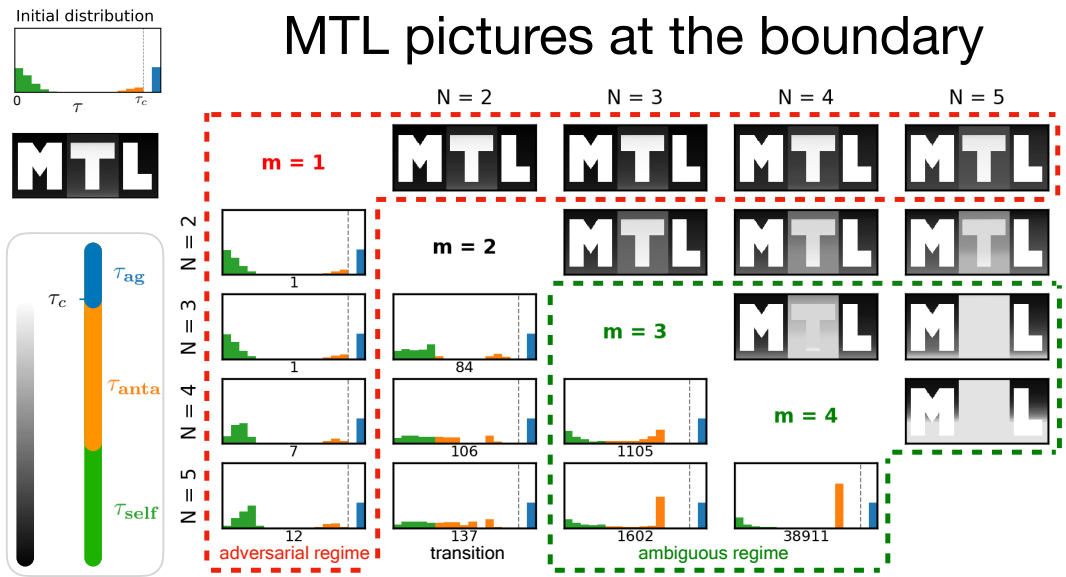
$$\frac{\partial T_{N,m}}{\partial \tau_b} = \frac{N\tau_b^{N-1} L_b - mT_{N,m}\tau_b^{m-1} L_b}{\sum_{i=1}^{M} \tau_i^m L_i} \tag{16}$$

where L_b is the number of ligands in the bth bin. We subtract this value multiplied by a small number ϵ from the exact binding times, as in Eq. 6 in the main text, and we compute a new output $T_{N,m}$. We repeat this procedure until $T_{N,m}$ dips just below the response threshold τ_c^{N-m}. We then display the ligand distributions. We bin ligands and compute the gradient in batches to prevent the gradient from becoming negligibly small. If we would compute the gradient for each ligand with an individual binding time, there would be exactly one ligand with that specific binding time, and because the gradient scales with L, we would need to go through many more iterations. Decreasing the binsize and step size ϵ may enhance the resolution, but is not required. We found good results by considering bins with a binsize of 0.2s and $\epsilon = 0.2$.

MTL pictures

We can visually recast immune recognition as an image recognition problem by placing pixels on a grid and coloring them based on their binding time with a given scale. We chose to let white pixels correspond to not self ($\tau > \tau_c$), gray pixels to antagonist ligands ($\tau_a < \tau < \tau_c$) and black pixels to self ligands $\tau \ll \tau_a$. We are free to introduce any kind of spatial correlation to create "immune pictures" from a ligand distribution. This results in what we term "MTL-pictures", Fig. 1. The initial ligand distribution, MTL picture and scale are given on the left. We perform iterative gradient descent like in the main text, and plot the ligand distribution and the corresponding immune pictures at the boundary for various (N, m) (Fig. 1). The results are striking. For a T cell operating in the adversarial regime, the "signal" MTL is unaltered at the decision boundary. At the transition $m = 2$, we see a slight change of color, while in the ambiguous regime, the signal actually changes from MTL to ML. As we desire for a robust decision-maker, the response should switch when the signal becomes significantly different. From this we conclude, *only in the robust regime can Montreal turn fully into the city of Machine Learning.*

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Appendix 3 Figure 1. MTL pictures. Explanation is found in the text

For the ligand distribution in the main text in Fig. 3 A, we have drawn $L_{\text{self}} = 7000$ from $\tau_{\text{self}} \in |\mathcal{N}(0, 0.33)|$ and $L_{\text{ag}} = 3000$ from $\tau_{\text{ag}} \in \mathcal{N}(3.5, 0.1)$. For the MTL pictures in Fig. 1, we have distributed the pixels in the 179x431 frame – equal to R, the number of receptors – as $L_{\text{self}} = 0.60R$, $L_{\text{a}} = 0.12R$, $L_{\text{ag}} = 0.28R$. We sampled self ligands from $\tau_{\text{self}} \in |\mathcal{N}(0, 0.33)|$, antagonists from $\tau_{\text{a}} \in \tau_c - |\mathcal{N}(0, 0.33)|$ and agonists from $\tau_{\text{ag}} \in |\mathcal{N}(3.5, 0.01)|$, and set $\tau_c = 3$. The picture is engineered such that the agonist ligands fill the M and the L, the antagonists fill the T (which is why the T is slightly darker than the M and L). The self ligands fill the area around the letters M, T and L, such that the self with highest binding time surround the T. We have chosen this example to make the effect of proofreading explicit (and of course because we are based in Montreal and study Machine Learning). This result is generic, and the ambiguity of instances at the decision boundary of a robust model can be visualized with any well-designed image. Scripts to reproduce Fig. 3 A and Fig. 1 are available at https://github.com/tjrademaker/advxs-antagonism-figs/.

Behaviour for small binding times

Consider a mixture with L_1 ligands at $\tau_1 > \tau_c$ and L_2 ligands with small binding time $\tau_2 \to \tau_\epsilon = \epsilon\tau_1 \ll \tau_1$. To understand the behaviour of $T_{N,m}$ as a function of τ_ϵ we expand $T_{N,m}$ in small variable $\epsilon = \frac{\tau_\epsilon}{\tau_1}$ as

$$\begin{aligned} T_{N,m}(\{L_1, \tau_1; L_2, \tau_\epsilon\}) &= \frac{\tau_1^N L_1 + \tau_\epsilon^N L_2}{\tau_1^m L_1 + \tau_\epsilon^m L_2} \\ &= \frac{1 + \epsilon^N \frac{L_2}{L_1}}{1 + \epsilon^m \frac{L_2}{L_1}} \tau_1^{N-m} \\ &\simeq \left(1 + \epsilon^N \frac{L_2}{L_1}\right)\left(1 - \epsilon^m \frac{L_2}{L_1}\right) \tau_1^{N-m} \\ &\simeq \tau_1^{N-m} - \tau_1^{N-m} \frac{L_2}{L_1} \epsilon^m + O(\epsilon^N), \end{aligned}$$

which confirms that up to a constant $T_{N,m} \propto -\epsilon^m \propto -\tau_\epsilon^m$ for m large and $\tau_\epsilon \ll \tau_1$, as well as that

$$\frac{dT_{N,m}}{d\tau_\epsilon} \simeq -m\tau_1^{N-m-1} \frac{L_2}{L_1} \epsilon^{m-1} \propto -\tau_\epsilon^{m-1}. \tag{17}$$

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix 4

Few-pixel attack

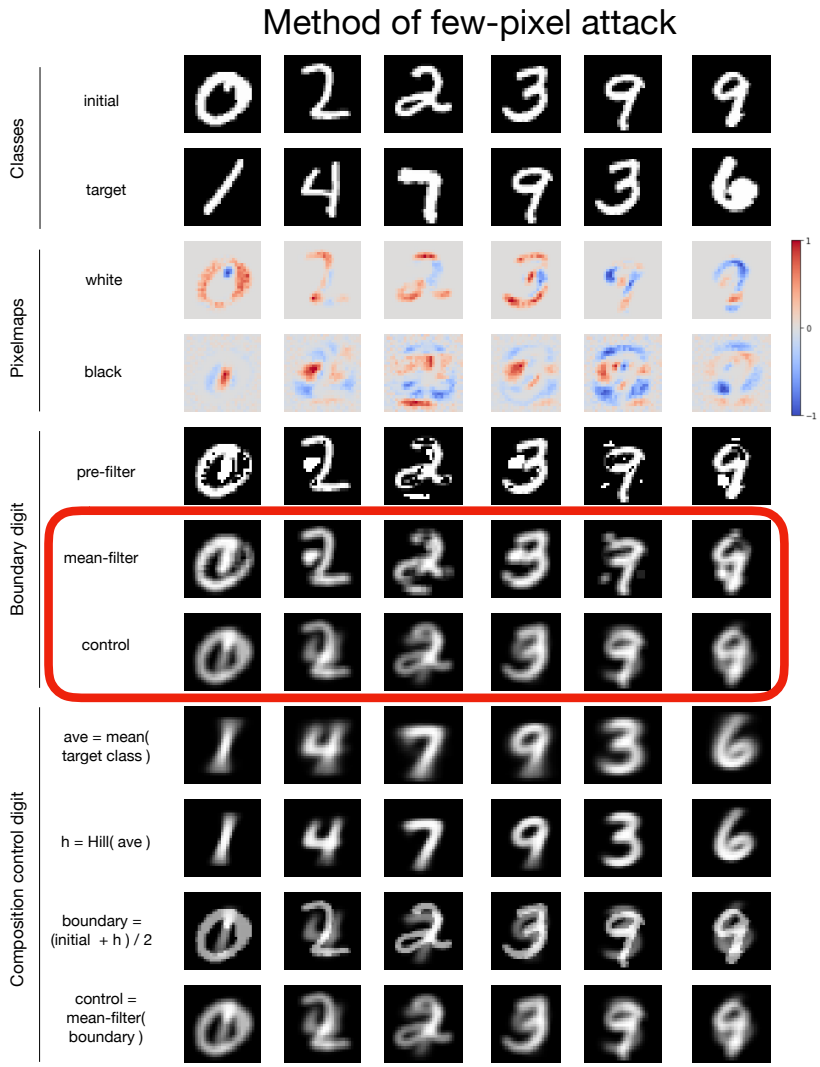
The few-pixel attack connects to ligand antagonism in the sense that few-pixels are needed to cause misclassification, corresponding to the addition of few maximally antagonizing ligands to a mixture fooling robust adaptive proofreading models. It is not the most efficient attack against a classifier without biomimetic defence, but it is the most efficient attack against classifiers with biomimetic defence, equivalent to adaptive proofreading models with $m > 1$. For these adaptive proofreading models, there exists a unique maximally antagonistic binding time, defined as the binding time that maximally reduces $T_{N,m}$.

With this in mind, we decided to make pixels black or white in a controlled manner, until the neural network classifies the perturbed, initial digit as the target class. In the following, we will refer to several stages of the few-pixel attack using Fig. 1. We first computed what we term pixelmaps. Pixelmaps contain the change of score when making a pixel white or black. In Fig. 1, blue colors correspond to pixels that will lower the score when turned white or black, while red colors are for pixels that will increase the score for the same operation. A grey color means the score is unchanged when whitening or blacking the pixel. The pixelmaps are scaled to the maximum change in score. We proceed in merging and sorting the pixelmaps from maximum to minimum change in score towards the target class, iteratively following the sorted list to decide which pixels in our digit to turn white or black. We do this until we reach the decision boundary (first iteration in which the digit is misclassified). The pre-filtered digits (Fig. 1, red rectangle) are the resulting boundary digits. They already contain perturbations corresponding to real features, but have an air of artificiality to them which allows us to fairly easily distill the ground truth. We remove this with a mean filtering [52], which is a 3x3 convolutional block that computes mean pixel values as

$$y_{i,j} = \frac{1}{9} \sum_{k,l=-1}^{1} x_{i+k,j+l}. \tag{18}$$

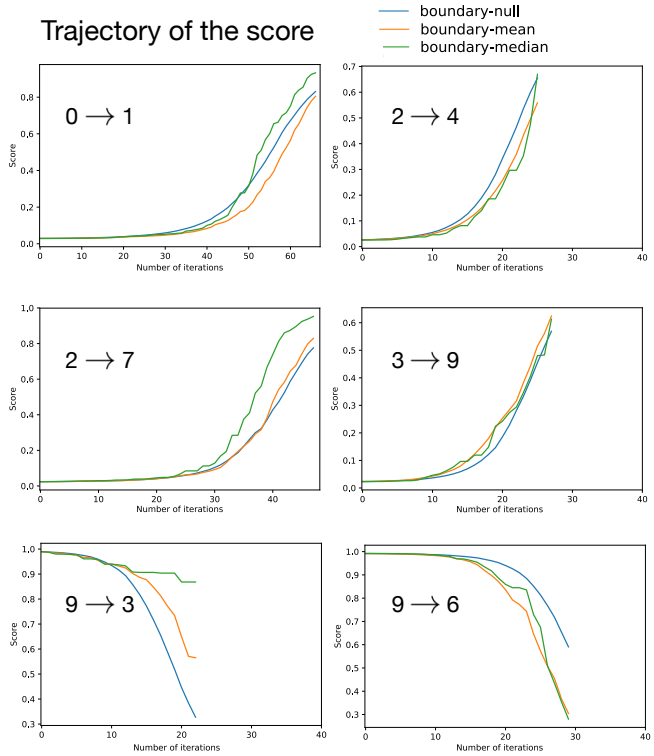
Biologically, this is pure receptor clustering, where a perturbation to a single receptor locally affects other ligands. Such digits are truly ambiguous digits that are tough to classify even as humans. These are the type of digits we expect to find on the decision boundary. Finally, we compare the mean-filtered digit at the decision boundary to the control: the sum of the initial digit and the hill function ($N = 3; \theta = 0.5$) on the average of all digits in the target class, then mean-filtered (Fig. 1 for a step-by-step composition). We apply the mean-filter to the control to again remove the artificiality of a digit plus an average, and make the comparison between boundary digit and control digit fairer. The similarity between mean-filtered boundary digit and control digit confirms our intuition that we are actually operating in the space between both classes when misclassification occurs.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Appendix 4 Figure 1. Method of few-pixel attack. Each column show how a few-pixel attack causes misclassification of an initial digit to a target class. The important result are the pre-filtered boundary digits and the control in the red rectangle. Pixelmaps determine which pixels increase (red) or decrease (blue) the score when turning an individual pixel in the initial digit white or black. We merge the pixelmaps, sort this list of pixels, and go through it from maximum to minimum change in score until misclassification occurs, resulting in the pre-filtered digit. We apply a mean-filter to make them look more like real digits, and indeed, these mean-filtered boundary digits closely resemble our control digits at the boundary. The control digits are composed of the mean-filtered initial digit plus locally contrasted (with hill function ($N = 3; \theta = 0.5$) average digit of the target class.

bioRxiv preprint doi: https://doi.org/10.1101/366724; this version posted March 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Appendix 4 Figure 2. Trajectory of the scoring functions of the attacks in Fig. S3. The blue, orange and green line correspond to various digits (actual digit, mean-filtered digit, median-filtered digit) for which we check the score, and terminate when reaching the boundary. The trajectory of the score for the null digit and the mean-filtered digit is generally the same. Moreover, the behavior of the score looks similar to the behavior of $T_{N,m}$ upon addition of maximally antagonizing ligands to a mixture of only agonist ligands in Fig. 2D in the main text.

We can also apply the mean-filter to the initial digit before generating the pixelmaps, and during the procedure, check the score on the mean-filtered perturbed image. This gives similar results, as we see by following the trajectory of the score for *boundary-null* and *boundary-mean*. We have shown the score explicitly in Fig. 2 for the digits in Fig. 1. The behavior of the score is remarkably similar to the interpolation between ligand mixtures (Fig. 2F, bottom panel in the main text). A nonlinear filtering method proposed in [52] is the median-filter, but this one works less well for black-and-white pixels.

We have shown examples that are generated when we select for instances where the number of iterations is large enough (20 suffices, we still consider this to be a few-pixel attack, keeping in mind that digits have 784 individual pixels). The authors of [43] specifically searched for single pixel attacks. Examples of single-pixel misclassification exist in our neural networks trained on two types of digits in MNIST too, but these we found non-informative. In cellular decision-making, this case corresponds to adding a single antagonist ligand to a ligand mixture to cause misclassification. This is only possible if the ligand mixture is already very close to the boundary. For such samples, we do not expect ambiguity to appear. Remember that near the boundary, the score landscape is steep, and small additions have a large effect.