Fooling the classifier: Ligand antagonism and adversarial examples

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 so-called adversarial perturbations. This is reminiscent of cellular decision-making where antagonist ligands may prevent correct signaling, like during the early immune response. We draw a formal analogy between neural networks used in machine learning and the general class of adaptive proofreading networks. We then apply simple adversarial strategies from machine learning to models of ligand discrimination. We show how kinetic proofreading leads to "boundary tilting" and identify three types of perturbation (adversarial, non adversarial and ambiguous). We then use a gradient-descent approach to compare different adaptive proofreading models, and we reveal the existence of two qualitatively different regimes characterized by the presence or absence of a critical point. These regimes are reminiscent of the "feature-to-prototype" transition identified in machine learning, corresponding to two strategies in ligand antagonism (broad vs. specialized). Overall, our work connects evolved cellular decision-making to classification in machine learning, showing that behaviours close to the decision boundary can be understood through the same mechanisms.

Introduction

Machine learning is becoming increasingly popular with major advances coming from deep neural networks [22]. Deep learning has improved the state-of-the-art in automated tasks like image processing [16], speech recognition [12] and machine translation [28], and has already seen a wide range of applications in research and industry. Despite their high-performance, neural networks suffer from blind spots: small perturbations added to unambiguous samples may lead to misclassification [29]. 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 [11]. Adversarial examples can be universal, and are often transferable across multiple architectures (see Akhtar and Mian [1] for a recent review). Two important properties should be noted: adversarial examples appear even in linear problems, and they can be significant notwithstanding the small L_∞ norm of the perturbation.

Another broad class of complex systems dealing with categorization and inference is found in cellular decision-making. For instance, immune cells have to discriminate between foreign and self ligands, irrespective of ligand quantity [5]. Such immune decisions are also prone to detrimental perturbations in a phenomenon called ligand antagonism [6]. Antagonism appears to be a general (and possibly necessary [7]) feature of cellular decision-making, and has been observed in T cells [2], mast cells [31] and other recognition processes like olfactory sensing [26].

In this work, we draw a correspondence between biophysical models of cellular decision-making displaying antagonism on the one hand, and adversarial examples in machine learning on the other hand. We start by drawing a formal analogy between feedforward neural networks and adaptive proofreading (or adaptive sorting) models [8, 20, 6], and we illustrate visually how to recast immune recognition into an image recognition problem. Then we show how direct adversarial perturbations correspond to antagonism by many weakly interacting ligands for the simplest model of adaptive proofreading. We refine our analysis by showing how kinetic proofreading steps work by "tilting" the decision boundary of cellular decision, corresponding to a strategy proposed in machine learning to defend against adversarial perturbations. We finally explore the geometry of the decision boundary for immune recognition, and we exhibit the emergence of a critical point, which we associate to a "feature-to-prototype" transition recently proposed in machine learning [18]

Results

Neural network for immune decision-making

We consider cellular decision-making based on ligand quality (notation τ) irrespective of quantity (notation L). An example can be found in immune recognition with the "lifetime dogma" [5], where it is assumed that a T cell discriminates ligands based on their characteristic binding time τ to T cell receptors. Defining τ_c as the activation threshold binding time, the problem boils down to ignoring many subthreshold ligands ($\tau < \tau_c$) while being able to respond to few agonist ligands with $\tau > \tau_c$ [2, 5, 8]. Ligand discrimination becomes a nontrivial problem when the cell cannot measure single-binding events, but only has access to global quantities such as the total number of bound receptors (Fig. 1 A).

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 [23, 14]). Those states allow the receptor to effectively compute different quantities, such as $c_m^i = k_i^{\text{on}} \tau_i^m, 0 \leq m \leq N$, which can be done with kinetic proofreading [23]. 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 [27, 21, 25], but on the time-scale of primary decision it is reasonable to assume that the ligand distribution does not change much [2].

A neural network-like diagram is displayed in Fig. 1 B to illustrate a general state-based decision-making mechanism based on those principles. The input nodes are the receptors, and the first layer of the network corresponds to the nodes C_m, summing up contributions of all receptors at a given proofreading step. With logarithmic activation functions between the first and second layer, the second layer effectively integrates these C_m to perform decision-making. While this network represents a simplified view of decision-making, it describes well the general class of "adaptive sorting" or "adaptive proofreading" models, accounting for many aspects of immune recognition [6]. Probability of decision-making in this context is a monotonically increasing function of the quantity

$$T_{N,m} = \frac{\sum_i k_i^{\text{on}} \tau_i^N}{\sum_i k_i^{\text{on}} \tau_i^m}. \tag{1}$$

If L ligands with identical τ and k^{on} are presented to the T cell, we find $T_{N,m} = \frac{k^{\text{on}} L \tau^N}{k^{\text{on}} L \tau^m} = \tau^{N-m}$, so that the binding times of the ligands can be directly evaluated irrespective of their quantity, corresponding to the lifetime dogma [9, 5]. If we now add L_{anta} antagonists with lower binding time $\tau_{\text{anta}} < \tau$ and equal on-rate k^{on}, we find $T_{N,m} = \frac{L\tau^N + L_{\text{anta}} \tau_{\text{anta}}^N}{L\tau^m + L_{\text{anta}} \tau_{\text{anta}}^m}$, which is smaller than the response τ^{N-m} for a single type of ligands, corresponding to ligand antagonism [10, 3, 2, 6] (Fig. 1 C). Fig. 1 D shows experimental curves of immune detection and antagonism (redrawn from [8]) compared to a model similar to Eq. 1 with $(N, m) = (4, 2)$ and identical on-rates[1]. To further the analogy with machine learning, we represent ligand mixtures as grayscale images in Fig. 1 E, where the intensity of each pixel corresponds to the binding time of a single ligand. The goal of the T cell is to detect the presence of white pixels, and because the T cell can not single out individual pixels, one thus needs to compute functions such as Eq. 1 to perform decision-making. The response is modulated by the binding times of "background" ligands. With our convention, it is easier for the system to detect white pixels in black background rather than in a gray background, corresponding to the idea that ligands close to threshold – more white than black – antagonize more strongly than ligands far from it, as proposed in [2]. This effect might be important for filtering out ambiguous cases [21], where many less white pixels could be mistaken for white ones.

Fast Gradient Sign Method and antagonism

Coming back to Eq. 1, $(N, m) = (1, 0)$ corresponds to a recently proposed model for antagonism in olfaction, with the role of k^{on} played by inverse affinity κ^{-1}, the role of τ played by efficiency η, and where the spiking rate of olfactory receptor neurons is a function $J(T_{N,m})$ [26], that can be interpreted as a scoring function in the machine learning sense. Notice that in this case, $T_{1,0}$ computes the average τ weighted by k_i^{on}.

[1] Offset for pure ligands is due to the addition of a small term in the denominator of Eq. 1 corresponding to a minimum ligand concentration for response, as expected from biology

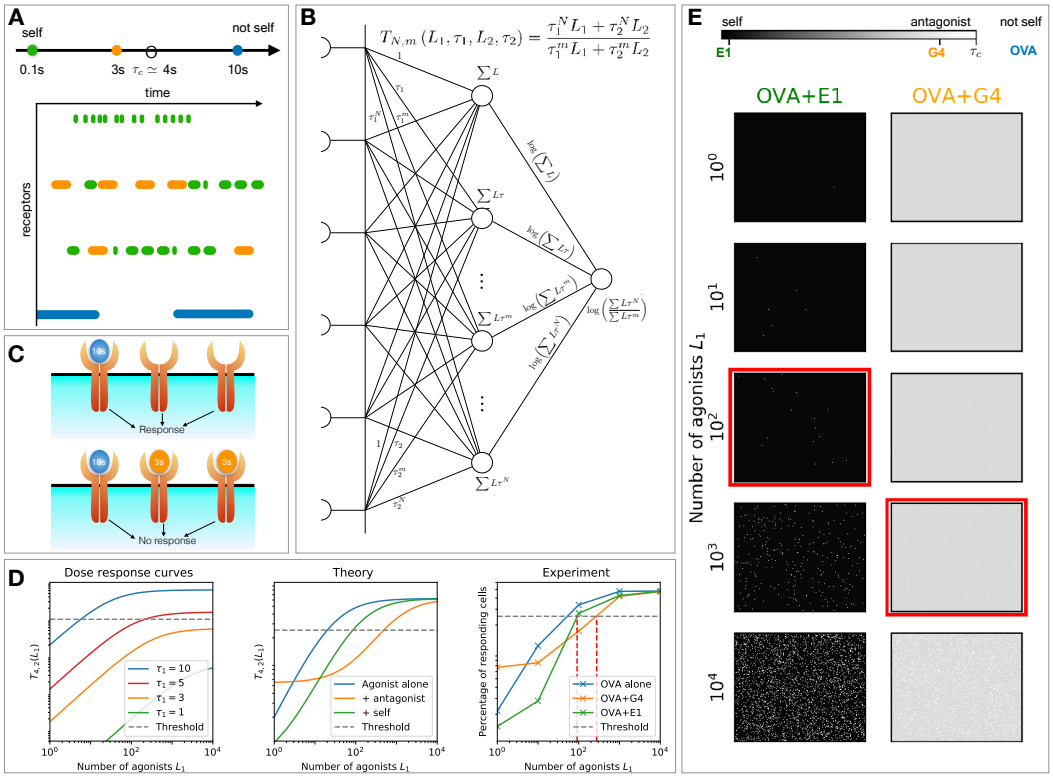


Figure 1. Ligand discrimination interpreted as an image recognition problem. A) Typical receptor occupancy of ligand binding events through time. B) Representation of an immuneural network. C) The T cell responds to few agonists alone, while in the presence of antagonists, it fails to notice them. D) Dose-response curves for ligands with different binding times for pure ligand types, and a mixtures for theory and experiment (redrawn from [8]). E) Immune pictures: schematics of images the T cell observes. The red frames mark when upon adding agonists, the T cell starts responding, corresponding to the red dashed lines in D), right panel.

We follow the original Fast Gradient Sign Method (FGSM) [11] by computing the maximum adversarial perturbation $\eta = \epsilon\,\mathrm{sgn}\,(\nabla_x J)$ with $||\eta||_\infty \leq \epsilon$ and $\nabla_x J$ the gradient of the scoring function. In our case, $\nabla_x J$ with respect to parameters k_i^{on} and τ_i gives

$$\eta = \epsilon\,\mathrm{sgn}\begin{pmatrix}\partial_{\tau_i} J\\ \partial_{k_i^{\mathrm{on}}} J\end{pmatrix} = \epsilon\,\mathrm{sgn}\,(A)\,\mathrm{sgn}\begin{pmatrix}k_i^{\mathrm{on}}\\ \tau_i - T_{1,0}\end{pmatrix}, \tag{2}$$

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 intuitive rules (Fig. 2 A).

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

Next, we consider important limiting cases. For instance, starting from L identical ligands with $\{k_{\mathrm{on}} = 1, \tau\}$, response $T_{1,0}^{\mathrm{before}} = \tau$ where τ itself is of order 1 (in proper units), a drastic change of response occurs if we suddenly add R ligands with short binding times τ_ϵ and small k^{on} of order ϵ, see Fig. 2 B[2]. We then have

$$T_{1,0}^{\mathrm{after}} = \frac{L\tau + \epsilon R \tau_\epsilon}{L + \epsilon R} = \frac{\tau + \frac{\epsilon R}{L}\tau_\epsilon}{1 + \frac{\epsilon R}{L}} \tag{3}$$

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

$$\frac{T_{1,0}^{\mathrm{after}} - T_{1,0}^{\mathrm{before}}}{T_{1,0}^{\mathrm{before}}} \simeq -\frac{\frac{\epsilon R}{L}}{1 + \frac{\epsilon R}{L}} \tag{4}$$

is of order 1 as soon as $\epsilon R \sim L$, giving a decrease comparable to the original response instead of being of order ϵ. The limit where ϵR is big thus corresponds to a strong antagonistic effect of many weakly bound ligands, which yields the same effect as "competitive antagonism" in olfaction [26] [3]. A similar change is observed if initially many ligands R have binding times τ between 0 and ϵ: decreasing their binding time to 0 yields a change of the numerator from $L\tau + R\epsilon \to L\tau$ while leaving the denominator unchanged to $L+R$, so that again the relative change $T_{1,0}^{\mathrm{after}}/T_{1,0}^{\mathrm{before}} - 1 \simeq -\frac{\epsilon R/L}{\tau + \epsilon R/L} = O(1)$ if $\epsilon R \sim L\tau$. Both these situations are reminiscent of the adversarial perturbations in [11] where it was observed that adding $\eta = \epsilon\,\mathrm{sgn}\,(\nabla_x J)$ leads to a significant perturbation on the scoring function J of order ϵN, with N being the (usually high) dimensionality of the input space (corresponding to the number of pixels). There is thus a direct correspondence between number of pixels in a picture and the high number of available receptors R. In both cases, the change of scoring function can be large despite the small amplitude ϵ of the perturbation, intuitively corresponding to a steep gradient in the maximally adversarial direction.

Boundary tilting and categorizing perturbations

In the immune context, such strong adversarial effects have to be mitigated because they correspond to antagonism by self [20]. This is done with kinetic proofreading [23, 2, 8], i.e. in our language by taking an output $T_{N,m}$ with $N > m > 0$. This ensures that self ligands with $\tau_\epsilon \ll 1$ barely contribute to $T_{N,m}$ since they appear in the numerator and denominator as τ_ϵ^N and τ_ϵ^m. Their contribution can be neglected compared to the contribution from other ligands with $\tau \sim 1$, even when the self ligands are numerous. This imposes an inverted hierarchy of antagonism, where the strongest antagonizing ligands exist closer to threshold [6], contrary to the case where $m = 0$. We now show that proofreading provides a boundary tilting effect, similar to what is described in machine learning [30] (Fig. 2 C, see Appendix 1 for an illustration of this effect on the discrimination of the original 3 vs 7 MNIST from Goodfellow et al. [11]).

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

$$T_{N,m}(\{L_1, \tau_1; L_2, \tau_2\}) = \frac{\tau_1^N L_1 + \tau_2^N L_2}{\tau_1^m L_1 + \tau_2^m L_2} \tag{5}$$

[2] Equivalently, we could have assumed that R short binding ligands change their k^{on} from 0 to ϵ

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

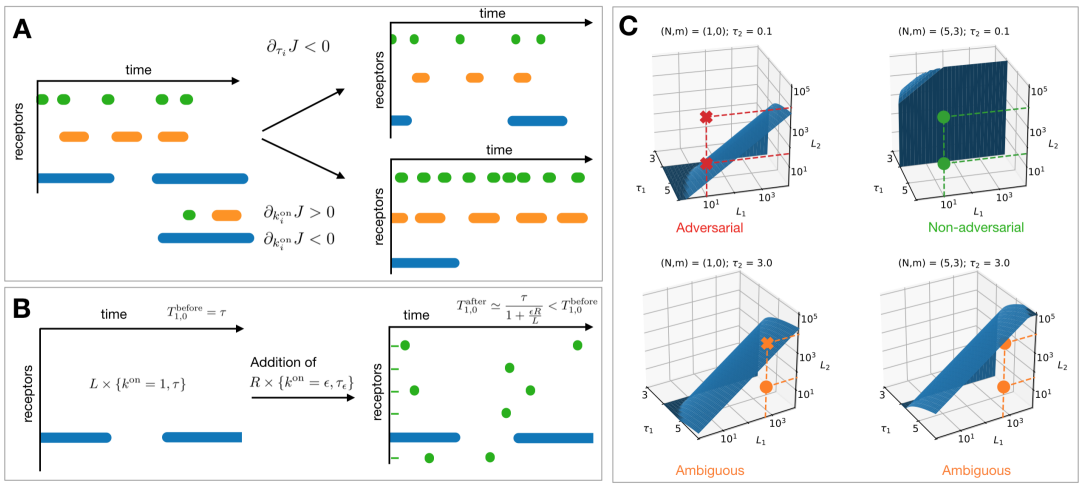


Figure 2. A) Three rules of ligand binding via FGSM. B) Change of response upon addition of R ligands with small τ_ϵ. C) Decision boundary of the immune model, where the expression of Eq. 5 is equal to τ_c^{N-m}. The region under the surface corresponds to cell response. The classifier with a single 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.

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 ($\mathcal{O}(\epsilon)$)

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. 2 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. 2 C, top left), 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 Appendix 1, [6]), 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. 2 C, top right) where the boundary is vertical, independent of L_2, such that decision-making is based only on L_1 ligands at τ_1 initially present . 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 $\tau_2 \simeq 0$ [6]. 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. 2 C, bottom). 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, while the gradient is weak, are of a different nature of the adversarial case of Fig. 2 C, top left, where the boundary is tilted as well but the gradient is steep. For this reason we call them "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.

Dichotomy of antagonism close to the decision boundary

These observations motivate a more precise study of the gradient towards the decision boundary. We follow Krotov and Hopfield, who studied a similar problem for MNIST digit classifier, encoded with generalized Rectified polynomials of variable degrees n [18] (reminiscent of the iterative FGSM introduced in Kurakin et al. [19]). The general idea is to find out how to most efficiently fool the system, and how this depends on the architecture of the decision algorithm. Krotov and Hopfield identified qualitative changes from a "feature" to a "prototype" encoding with increasing n, accompanied by a better resistance to adversarial perturbations [17, 18]. While for small n, digits on the boundary are the initial digits to which a weak, distributed perturbation is added (corresponding to the learned "features"), for big n, they are intermediate forms with no clear identity, as would be expected for ambiguous digits that are already difficult to recognize for human observers (Fig. 3 A).

We consider the dynamics of a ligand mixture when following the gradient of $T_{N,m}$, and study how to most efficiently fool the decision-maker (or in biological term, how to best antagonize it). We iteratively change the binding time of nonagonist ligands with $\tau < \tau_c$ to

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

while keeping the distribution of foreign ligands with $\tau > \tau_c$ constant. Biologically, these dynamics should 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 [15, 24] and associated vaccine failure [13].

We consider two initial ligand distributions with different visual representations: one with pixels randomly distributed (Fig. 3 C), the other with pixels arranged to form the letters "MTL" (Fig. 3 D). The letter "T" contains ligands (pixels) just below threshold while the "M" and "L" are made up of ligands above threshold. We then 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 boundary in the descent defined by Eq. 6. In both cases, for small m, we see strong adversarial effects, as the boundary is almost immediately reached. As m increases, in Fig. 3 B the distribution of ligands concentrate around one peak visually corresponding to a weak "whitening" of our visual representation, while the two peaks in Fig. 3 C approach each other. For $m = 2$, a qualitative change occurs: the ligands suddenly spread over a broad range of binding times in both Figs 3 B and C , 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 in the gray "antagonistic" zone. Visually, this corresponds to black pixels reappearing for higher m while all other pixels turn white gray, which gives pictures at the boundary very different from the original ones. Consistent with this, the number of ϵ-sized steps to reach the boundary is 3 to 4 orders of magnitude higher for $m > 2$ than for $m < 2$. The qualitative difference is most striking for the "MTL" case: for $m \leq 2$, one can still distinguish the three letters while for $m > 2$, the "T" almost entirely disappears so that "MTL" (Montreal) turns into (the city of) "ML" (Machine Learning).

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 $\tau_1 > \tau_c$ and self $\tau_2 = \tau_\epsilon$, similar to equation 5), 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 2 for this calculation). In particular, for $0 < m < 1$, $\frac{\partial T_{N,m}}{\partial \tau_\epsilon} \propto -\tau_\epsilon^{m-1}$ diverges as $\tau_\epsilon \to 0$, which corresponds to the steep gradient described above for adversarial perturbations. In this regime, the ligands close to $\tau_\epsilon \sim 0$ follow the steep gradient to quickly localize close to the minimum of this potential (unimodal distribution of ligand for small m on Fig. 3 B, C). 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 $\tau_\epsilon = 0$, and an inflexion point (square) appears in between the minimum (circle) and $\tau_\epsilon = 0$ (Fig 3 D). This explains the sudden broadening of the ligand distribution, and the associated increase in the number of steps to reach the decision boundary. For $m > 2$, ligands close to 0 are pinned while only ligands with large enough $\frac{\partial T_{N,m}}{\partial \tau}$ can efficiently move towards the minimum (which is closer to the boundary as N, m increases). Flatter potentials are obtained for large N, m, which explains the many required iterations to reach the boundary.

The change at $m > 2$ is strongly reminiscent of the transition observed by Krotov and Hopfield in their study of gradient dynamics similar to eq. 6 for rectified polynomials with increasing degree n [18], applied on digit classifiers. This is best visible in Fig. 3 B. For $m < 2$, all "background" ligands below threshold only slightly change their τ, corresponding to a broad non-specific antagonizing effect, reminiscent of the speckled pattern for low n in Fig. 3 A. The distribution of ligands in Fig. 3 B then barely changes and stays unimodal. Conversely, for $m > 2$, the main antagonizing effect comes from

a specialized subpopulation of ligands, corresponding to the appearance of a bimodal distribution where some ligands "localize" at the maximally antagonizing τ just below threshold (minimum of $T_{N,m}$ in Fig. 3. D). Similarly, only the subset of pixels that spatially correlate with the initial digit change value in Fig. 3 A for big n. In Fig. 3 C, the same dichotomy between global and specialized antagonism is observed. For $m < 2$ the pictures barely change, indicating a non-specific antagonizing effect, while for $m > 2$, only the part corresponding to the "T" part of the pictures changes while the black background of the "M" and "L" remains unchanged, which is again reminiscent of the ambiguous digits on Fig. 3 A for big n.

Discussion

Using gradient-based methods from machine learning we can phenomenologically relate adversarial examples to ligand antagonism. Simple models are fooled by gradient-based methods and mitigate the effects by tilting the decision boundary with kinetic proofreading. Gradient descent close to the boundary can display two qualitative behaviours reminiscent of what is observed in machine learning, which we further characterize by the appearance of a critical point for $m \geq 2$ for ligands at $\tau = 0$. Interestingly, the models of adaptive proofreading presented here were first generated with *in silico* evolution [20]. Strong antagonism naturally appeared in the simplest simulations and required modification of objective functions very similar to what has been done at the same time for adversarial examples in machine learning [11]. Both ligand antagonism and adversarial examples appear to be instances of the general phenomenon of fooling the classifier. Adaptive proofreading models as presented here are arguably the simplest instance of this phenomenon, amenable to analytical studies and helpful to build our intuition to perturbations in decision-making.

A caveat of our approach is that in biophysical models a clear decision axis in the τ direction exists, which is not usually the case in machine learning. Here the algorithm has to effectively learn representations, such as pixel statistics and spatial correlations in images [16]. Case in point, a spatial transformation was recently proposed as an adversarial attack to exploit such adversarial directions [33]. However, underlying, low manifold descriptions could still combine higher level information in ways similar to our individual parameters τ so that the theory presented here could still apply once those directions are discovered.

Many internal biological systems have evolved to perform decision-making, and it is fascinating that quantitative studies of such systems allow for connections with machine learning. From the biology standpoint, it means that deep insights might 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 changing agents can present themselves in two distinct regimes when confronted with cellular detectors. Biochemically, antagonism manifests itself via broadly distributed antagonizing ligands ($m < 2$) or via specific optimization of ambiguous antagonists ($m > 2$), depending on the detection mechanism used by immune cells. From the defence standpoint, the case $m \geq 2$ appears to be much more resistant to adversarial perturbations, and thus would be most relevant in an immune context where detectors (immune cells) have to filter out antagonistic perturbations. This might be relevant for the pathology of HIV infections [15, 24, 13] or, more generally, could provide explanations on the diversity of altered peptide ligands [32]. The case $m < 2$ with a steep gradient might be more relevant in signaling contexts, where it might be valuable to separate well mixtures of inputs. For olfaction it has been suggested that such 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 [26].

On the machine learning side, new inspirations coming from biology are not restricted to classical sensory systems or neuroscience, but may also come from cellular decision-making, such as immune recognition. Our results in Fig. 3 suggests that flattening of the adversarial directions via a critical point might be key to resisting adversarial perturbations, yielding qualitative changes in the dynamics towards the decision boundary. Those are characterized by a subpopulation of inputs moving towards the boundary to define samples ambiguous to a human observer. This is reminiscent of the perturbed animal pictures fooling humans [4] e.g. with chimeric images that combine different animal parts (such as spider and snake). Lastly, one can show mathematically in the biophysical context that it is not possible to fully get rid of antagonism close to the boundary [6]. Bypassing this requires that the "signal" of the agonist ligands is strong and far enough removed from the boundary (which can be done using many proofreading steps). Similar inevitability theorems might be generalizable to machine learning.

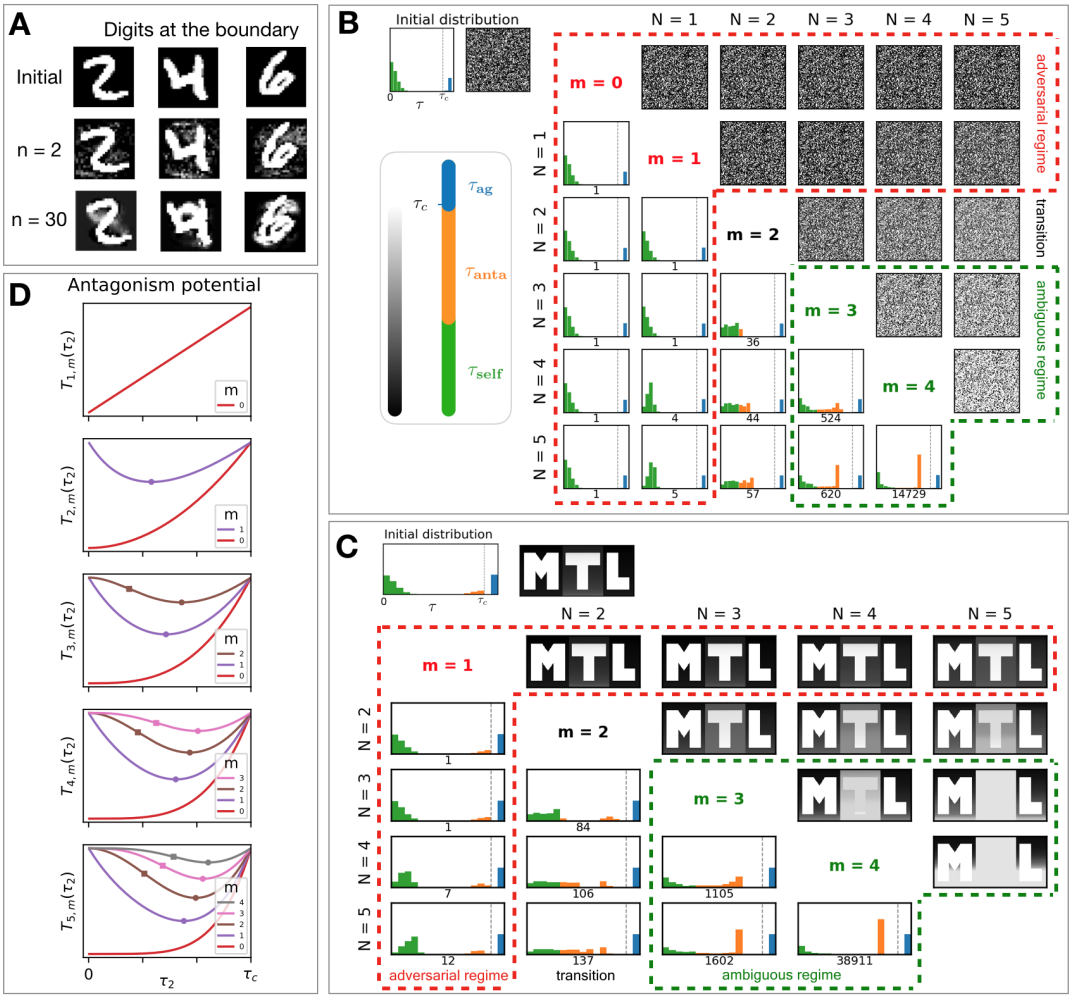


Figure 3. Characterization of the decision boundary. A) Steepest descent towards the decision boundary for a specialized Hopfield network with linear ($n = 2$) and highly nonlinear ($n = 30$) activation functions. Reproduced from [18]. B) Ligand distribution at the boundary. For various cases (N, m) we change the binding time of self ligands of an initial distribution along the steepest gradient until reaching the decision boundary, after which we draw the resulting immune picture (upper triangular) and ligand distribution (lower triangular). The number of steps needed in the gradient descent to reach the boundary is indicated below the distributions. C) Immune pictures with spatial correlations, evolved similar to B. D) $T_{N,m}(\{L_1, \tau_1; L_2, \tau_2)\}$ as a function of τ_2 for various (N, m). Antagonism strength of nonagonists is maximal when $T_{N,m}$ is minimal. Minima and inflexion points are indicated with a circle and square.

Acknowledgement
We thank Joelle Pineau and member of the François group for useful discussions.

References
[1] Akhtar, N. and Mian, A. (2018).

[2] Altan-Bonnet, G. and Germain, R. N. (2005). Modeling T cell antigen discrimination based on feedback control of digital ERK responses. *PLoS Biology*, 3(11):1925–1938.

[3] Dittel, B. N., Stefanova, I., Germain, R. N., and Janeway, C. A. (1999). Cross-antagonism of a T cell clone expressing two distinct T cell receptors. *Immunity*, 11(3):289–298.

[4] Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. (2018). Adversarial Examples that Fool both Human and Computer Vision.

[5] Feinerman, O., Germain, R. N., and Altan-Bonnet, G. (2008). Quantitative challenges in understanding ligand discrimination by $\alpha\beta$ T cells. *Molecular Immunology*, 45(3):619–631.

[6] François, P. and Altan-Bonnet, G. (2016). The Case for Absolute Ligand Discrimination: Modeling Information Processing and Decision by Immune T Cells. *Journal of Statistical Physics*, 162(5):1130–1152.

[7] François, P., Hemery, M., Johnson, K. A., and Saunders, L. N. (2015). Phenotypic spandrel: absolute discrimination and ligand antagonism. *Physical Biology*, 13(6).

[8] François, P., Voisinne, G., Siggia, E. D., Altan-bonnet, G., and Vergassola, M. (2013). Phenotypic model for early T-cell activation displaying sensitivity , specificity , and antagonism. *Proceedings of the National Academy of Sciences of the United States of America*, 110(10):E888—-E897.

[9] Gascoigne, N. R., Zal, T., and Alam, S. M. (2001). T-cell receptor binding kinetics in T-cell development and activation. *Expert Reviews in Molecular Medicine*, 2001(06):1–17.

[10] Germain, R. N. and Stefanova, I. (1999). The Dynamics of T Cell Receptor Signaling: Complex Orchestration and the Key Roles of Tempo and Cooperation. *Annual Review of Immunology*, 17(1):467–522.

[11] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples.

[12] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. (November).

[13] Kent, S. J., Greenberg, P. D., Hoffman, M. C., Akridge, R. E., and McElrath, M. J. (1997). Antagonism of vaccine-induced HIV-1-specific CD4+ T cells by primary HIV-1 infection: potential mechanism of vaccine failure. *Journal of Immunology*, 158(2):807–815.

[14] Kersh, G. J., Kersh, E. N., Fremont, D. H., and Allen, P. M. (1998). High- and low-potency ligands with similar affinities for the TCR: The importance of kinetic in TCR signaling. *Immunity*, 9(6):817–826.

[15] Klenerman, P., Rowland-Jones, S., McAdam, S., Edwards, J., Daenke, S., Lalloo, D., Köppe, B., Rosenberg, W., Boyd, D., and Edwards, A. (1994). Cytotoxic T-cell activity antagonized by naturally occurring HIV-1 Gag variants. *Nature*, 369(6479):403–407.

[16] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks.

[17] Krotov, D. and Hopfield, J. J. (2016). Dense Associative Memory for Pattern Recognition.

[18] Krotov, D. and Hopfield, J. J. (2017). Dense Associative Memory is Robust to Adversarial Inputs.

[19] Kurakin, A., Goodfellow, I., and Bengio, S. (2016). Adversarial Machine Learning at Scale.

[20] Lalanne, J.-B. and François, P. (2013). Principles of adaptive sorting revealed by in silico evolution. *Physical Review Letters*, 110(21):1–5.

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

[22] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

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

[24] Meier, U. C., Klenerman, P., Griffin, P., James, W., Köppe, B., Larder, B., McMichael, A., and Phillips, R. (1995). Cytotoxic T lymphocyte lysis inhibited by viable HIV mutants. *Science*, 270(5240):1360–2.

[25] Mora, T. (2015). Physical Limit to Concentration Sensing Amid Spurious Ligands. *Physical Review Letters*, 115(3):038102.

[26] Reddy, G., Zak, J. D., Vergassola, M., and Murthy, V. N. (2018). Antagonism in olfactory receptor neurons and its implications for the perception of odor mixtures. *eLife*, 7:e34958.

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

[28] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks.

[29] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks.

[30] Tanay, T. and Griffin, L. (2016). A Boundary Tilting Perspective on the Phenomenon of Adversarial Examples.

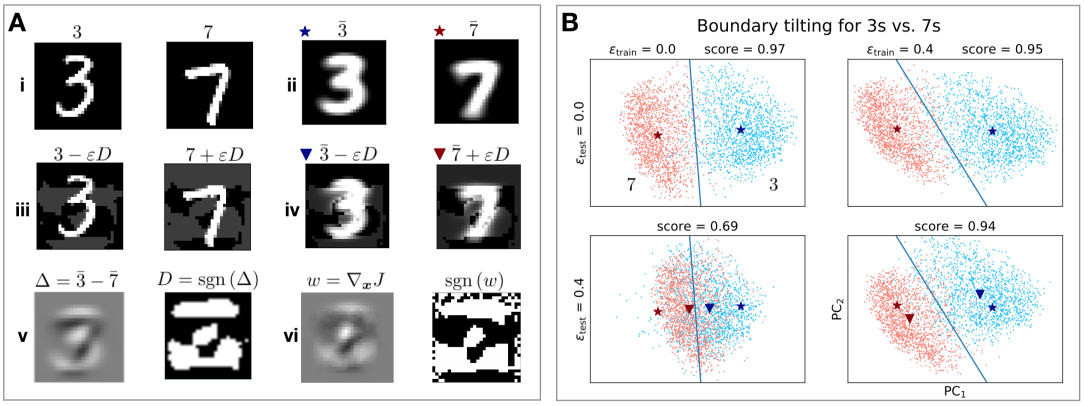
[31] Torigoe, C., Inman, J. K., and Metzger, H. (1998). An Unusual Mechanism for Ligand Antagonism. *Science*, 281(5376):568–572.

[32] Unanue, E. R. (2011). Altered Peptide Ligands Make Their Entry. *Journal of Immunology*, 186(1):7–8.

[33] Xiao, C., Zhu, J.-y., Li, B., He, W., Liu, M., and Song, D. (2018). Spatially transformed adversarial examples.

Appendix 1

Boundary tilting in digit classification



Appendix 1 Figure 1. Boundary tilting in digit classification. A) 3s and 7s. (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 closely to (vi), the perturbation found with FGSM B) Projection of 3s and 7s on its PCs. The classes are separated by the blue line from a linear Support Vector Machine, and the triangle and star show the average of the classes with and without adversarial perturbation. From (i) to (iv) we have cycled through permutations of perturbing training and/or test set with their specific adversarial perturbation. On the right panels, note how the boundary has tilted and the triangle moved away from the star parallel to the decision boundary.

The goal of this section is to illustrate on a very simple example boundary tilting in a machine learning context, namely the MNIST dataset as explored first in [11], which happens to be almost linear.

We are interested in a binary classification problem (response vs. no response), thus, it suffices to take a subset of 3s and 7s from MNIST. Typical 3's and 7's are shown in Fig. 1 A (i-iv). Tanay and Griffin [30] pointed out that the adversarial perturbation generated with the Fast Gradient Sign Method (FGSM) proposed in [11] can also be found via $D = \mathrm{sgn}\,(\bar{3} - \bar{7})$, Fig. 1 A (v). Note its similarity to the FGSM adversarial perturbation $\mathrm{sgn}\,(w) = \mathrm{sgn}\,(\nabla_x J)$ (Fig. 1 A (vi)).

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 Support Vector Machine (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 accuracy of detection for single cells (e.g. immune cells present false negative rates of 10 % for strong antagonists [2]) The red and blue star in the figure denote the average digit $\bar{3}, \bar{7}$.

Next, we transformed the test set as $3 \to 3' = 3 - \epsilon_{\mathrm{test}} D, \quad 7 \to 7' = 7 + \epsilon_{\mathrm{test}} D$, where $\epsilon_{\mathrm{test}} = 0.4$ is the strength of the adversarial perturbation (Fig. 1 A (iii)). $\bar{3}'$ and $\bar{7}'$ moved towards each other 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 $\epsilon_{\mathrm{train}} D_{\mathrm{train}} = \epsilon_{\mathrm{train}}(\bar{3}_{\mathrm{train}} - \bar{7}_{\mathrm{train}})$ to the images in the training set, computing the new PCs and training the linear regression model. This effectively "tilts" the decision boundary, while keeping 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, thereby preserving the good classification accuracy, giving a simple example of the more general phenomenon studied in [30].

Scripts for boundary tilting in ligand discrimination and digit discrimination are available online at https://github.com/tjrademaker/advxs-antagonism-figs/.

Gradient in the L_2 direction

We recall results from [7] 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

$$T_{N,m}(L_1, \tau_c) = \tau_c^{N-m}. \tag{7}$$

The main result of [7] is the linear response of $T_{N,m}(L_1, \tau_c)$ to the addition of $\{L_2, \tau_c - \epsilon\}$ subthreshold ligands.

$$T_{N,m}\left(\{L_1, \tau_c; L_2, \tau_c - \epsilon\}\right) = T\left(L_1 + L_2, \tau_c\right) - \epsilon L_2 \mathcal{A}\left(L_1 + L_2, \tau_c\right) \tag{8}$$

$$= \tau_c^{N-m} - \epsilon \frac{L_2}{L_1 + L_2} \frac{d}{d\tau} T_{N,m}(L_1 + L_2, \tau)\Big|_{\tau=\tau_c}, \tag{9}$$

where we used the definition

$$\mathcal{A}\left(L, \tau_c\right) = \frac{1}{L} \frac{d}{d\tau} T_{N,m}(L, \tau)\Big|_{\tau=\tau_c}. \tag{10}$$

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

$$\frac{\tau_c - \tau_2}{L_1}. \tag{11}$$

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

$$\mathcal{A}(L_1, \tau_c) = \frac{1}{L_1} \frac{d}{d\tau} \tau\Big|_{\tau=\tau_c} = \frac{1}{L_1}. \tag{12}$$

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.

Appendix 2

Gradient descent towards the boundary

Our immune model is well-suited to characterize the decision boundary between two classes, because of the analytical classifier. We want to know how to most efficiently change the binding time of the short binding self to cause the immune model to reach the decision boundary. We have taken inspiration from [18] and adapted our approach from the iterative FGSM [19]. 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 $\tau_b < \tau_c$

$$\frac{\partial T_{N,m}}{\partial \tau_b} = \frac{N\tau_b^{N-1} L_b - mT_{N,m}\tau_b^{m-1} L_b}{\sum_{i=1}^{M} \tau_i^m L_i} \tag{13}$$

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}. Finally, we display the ligand distribution and the immune pictures, like we did in Fig. 3 in the main text. The reason why we bin ligands and compute the gradient in batches is 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 $\epsilon = 0.2$. Scripts to reproduce Fig 3 B-D are available online at https://github.com/tjrademaker/advxs-antagonism-figs/.

For the immune pictures at random ligand order in Fig. 3 B, we have drawn $L_\mathrm{self} = 7000$ from $\{\tau_\mathrm{s}\} \in |\mathcal{N}(0, 0.33)|$ and $L_\mathrm{ag} = 3000$ from $\{\tau_\mathrm{ag}\} \in \mathcal{N}(3.5, 0.1)$. For the MTL $\rightarrow$ ML transition in Fig. 3 C, we have distributed the pixels in the 179x431 frame – appropriately set equal to R, the number of receptors – as $L_\mathrm{self} = 0.60R$, $L_\mathrm{anta} = 0.12R$, $L_\mathrm{ag} = 0.28R$. We sampled self ligands from $\{\tau_\mathrm{s}\} \in |\mathcal{N}(0, 0.33)|$, antagonists from $\{\tau_\mathrm{anta}\} \in \tau_c - |\mathcal{N}(0, 0.33)|$ and agonists from $\{\tau_\mathrm{ag}\} \in |\mathcal{N}(3.5, 0.01)|$, 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 its color 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 the true decision boundary can be visualized with any well-designed image.

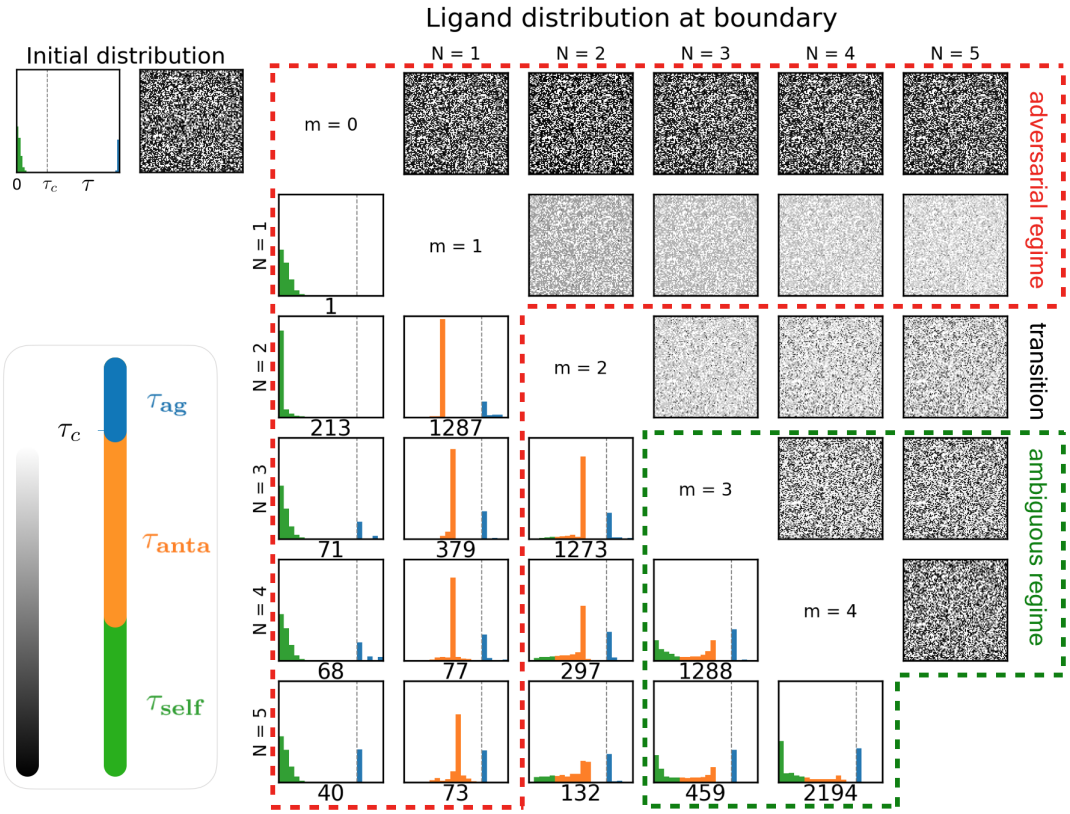
Effect of agonist binding times

For a given initial distribution, the distribution at the boundary is robust to parameter values, although neither mean nor variance of the agonist distribution can be too large. In that case, for high N and m, the decision boundary cannot be reached, even with all self ligands employed at the maximally antagonizing τ. A prediction from our model is that with sufficient L_ag at a high enough τ_ag a fixed number of self ligands L_self are not able to antagonize the response. This is a way to escape the inevitability theorem, which states that close to the boundary, there always exists an adversarial perturbation that causes misclassification, no matter how strong the adversarial defence is. Perceptibly changing the signal naturally allows for a change of classification, but these macroscopic perturbations are no longer small or imperceptible, and do not fall under the umbrella of adversarial examples.

If we allow for changes in $\{\tau_\mathrm{ag}\}$, we find that after the final iteration, $\{\tau_\mathrm{ag}\}$ has gone below τ_c, such that we are not capturing antagonistic effects anymore, instead we trivially find that we lose response when there is no more signal to respond to. We could have taken a lower limit for agonists, i.e. $\{\tau_\mathrm{ag}\} \geq \Theta$, but the value of Θ is arbitrary and not easily justifiable. For one, it should be larger than τ_c to still provide a net weight to the response class agonist ligands. In Fig. 1 we have computed the distribution at the boundary in the case where we allow agonists to change binding times until a threshold. We set $\Theta = 3.1$s to be the center of the first bin past τ_c, sample $L_\mathrm{self} = 6000$ from $\{\tau_\mathrm{s}\} \in |\mathcal{N}(0, 0.33)|$ and $L_\mathrm{ag} = 4000$ from $\{\tau_\mathrm{ag}\} \in 10 - \mathcal{N}(0, 0.1)$. Even at $m = 0$, we have to go through some iterations before reaching the boundary: we are sufficiently far away from the boundary due to the long binding agonists. The agonist ligands change binding times quickly because the gradient

$$\frac{\partial T}{\partial \tau_b} \propto \left(\frac{\tau_b}{\tau_c}\right)^{N-m-1} \tag{14}$$

is much steeper for agonists with $\tau_b > \tau_c$ than for nonagonists. This is also clear from the antagonism potential in Fig. 3 D in the main text: above τ_c the gradient gets only steeper. For $m \geq 1$, all agonists congregate in the first bin $\tau_b \sim \tau_c$. We observe an overall graying of the immune pictures at $m = 1$, which is undone at higher m when only a subset of nonagonist ligands changes binding time and antagonizes the response. This gives the typical bimodal distribution at larger (N, m), again due to the flat gradient at $\tau \sim 0$. It provides more evidence for the appearance of a critical point in a robust adversarial defence.



Appendix 2 Figure 1. Characterization of the decision boundary when agonists are not constant.

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

$$\frac{dT_{N,m}}{d\tau_\epsilon} \simeq -m\tau_1^{N-m-1} \frac{L_2}{L_1} \epsilon^{m-1} \propto -\tau_\epsilon^{m-1}. \tag{15}$$